EFFECTS OF AUDITORY FEEDBACK AND REAL-TIME VISUAL FEEDBACK ON SECOND LANGUAGE TONE LEARNING

BY

LI-HSIN NING

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
with a concentration in Second Language Acquisition and Teacher Education
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Chilin Shih, Chair and Co-Director of Research
Associate Professor Torrey Loucks, Director of Research
Associate Professor Tania Ionin
Assistant Professor Aaron Johnson
Associate Professor Ryan Shosted

ABSTRACT

Tone learning is challenging for non-tone speakers. While pitch, like melody, is processed by the right hemisphere for non-tonal language speakers, lexical tone as well as other language aspects (such as phonemes, syntax and semantics) appears to be processed by the left hemisphere for tonal language speakers. Adult second language (L2) learners of Mandarin whose native language is non-tonal may need to acquire left hemisphere dominance of pitch processing for successful tone learning. L2 learners of Mandarin have to acquire new perceptual categories for discriminating and identifying lexical pitch variation along with new sensorimotor skills to produce the rapid tone changes. Language learning, therefore, opens windows for observing changes in behavior and brain functions that impact the learner in multiple ways.

I addressed the neural plasticity indirectly by constructing a paradigm that incorporates auditory feedback called pitch-shift paradigm. The pitch-shift paradigm, in which a short and artificial change in pitch is fed back to speakers during vocalization, has been used to investigate how sensory information affects the way we control our speech motor activities. Pitch-shift responses (vocal responses to auditory perturbation) have reflex-like properties and are hard to suppress. So pitch-shift paradigm can be used to understand the stability of internal models for tone and to reveal how internalized pitch representations are built or reshaped.

Study 1 examined the language experience effect on pitch-shift responses and non-/linguistic tone discrimination. Discrimination of musical tones was correlated significantly with discrimination of Mandarin tones with the clearest advantage among Mandarin speakers and some advantage for L2 learners. Group differences were found in fundamental frequency (F0) contour shape in response to pitch-shift stimuli. F0 contours of Mandarin speakers were quantitatively least affected by the amplitude and direction of pitch perturbations, suggesting

more stable internal tone models. F0 contours of naïve speakers and L2 learners were significantly altered by the perturbation. The findings provide quantitative measurements on how language experience affects tone discrimination and voice F0 control (internal models), and establish the validity of using pitch-shift paradigm to document the success of tone learning.

Study 2 included trained vocalists as additional group to investigate the vocal training experience effect on audio-vocal responses and tone discrimination. Mandarin speakers performed significantly better on (adaptive) Mandarin tone discrimination compared to the other three groups. Mandarin speakers also showed more attenuation of pitch shift response amplitude during production of both the sustained vowel (nonlinguistic domain) and Mandarin tones (linguistic domain), especially compared to naïve speakers. The findings suggest Mandarin speakers have more robust pitch control over vocalization and are thus less affected by perturbation in auditory feedback. Trained vocalists also appear to rely more on internal models than naïve speakers in order to regulate voice F0 in the nonlinguistic (sustained vowel) domain, but not in the linguistic domain (Mandarin tone). L2 learners showed only subtle variation in production relative to the other groups.

Study 3 explored how enhancing feedback by adding real-time visual feedback to tone production could influence suprasegmental control critical for tones. Results show that both naïve speakers and Mandarin speakers can benefit from the use of real-time visual feedback for stabilizing their voice F0.

In this dissertation, I investigated how language experience, vocal training experience, and feedback mode may contribute to successful tone learning by using the pitch-shift paradigm. Native Mandarin speakers demonstrate robust internal models for lexical tones that are evident in perception and production, making language experience an important factor in shaping sensorimotor control. Trained vocalists' resemblance to Mandarin speakers shows that formal

vocal training may speed up tone learning. Furthermore, feedback in any format (auditory or visual) is important for language learning. Motor/feedforward commands for new speech sounds are built and strengthened by repeated practice with the help of auditory feedback. Visual feedback on voice accuracy could also facilitate the construction of language-specific feedforward commands for speech. All in all, the pitch-shift paradigm could be used effectively to tap into speakers' internal models of lexical tones. Robust internal models built through extensive language experience enable native speakers of Mandarin to maintain steady voice F0 in both linguistic and nonlinguistic domains, when their pitch is altered unexpectedly. Second language learning experience may have changed L2 learners' internal models because their vocal responses to pitch perturbation were not like naïve speakers'. Although the L2 learners have not acquired native-like internal models of lexical tones, it is expected that extensive exposure to Mandarin will reshape their sensorimotor control.

To my dear father and mother

ACKNOWLEDGMENTS

I cannot express enough thanks to my advisors, Dr. Torrey Loucks and Dr. Chilin Shih, for their continued support and encouragement. I would like to offer my sincere appreciation for the learning opportunities provided by my advisors. I also appreciate that they read the dissertation very carefully and spent hours with me in rehearsing the defense. Their valuable and constructive comments significantly improved the dissertation and made the defense very successful.

Under Dr. Loucks' supervision, I started to step into a field (neurophysiology) that I had not been familiar with. My technical skills for running pitch-shift experiments and doing analyses with MATLAB were trained by Dr. Loucks, who always saved me from trouble during the experiments. Dr. Loucks is a very helpful, knowledgeable and approachable professor. I appreciate that he always answered my questions right away. His instant responses made me feel very supported.

I would like to express my gratitude to Dr. Shih, who has been my academic advisor since the first year of my PhD. Dr. Shih is a far-sighted scholar and suggested new ways of analyzing the pitch-shift response data in my dissertation. I appreciate that she hired me as her Chinese TA, from which I learned a lot of computational skills that became my strength in job placement. Her continuous support throughout the years enabled me to complete my research. I also thank Dr. Shih for inviting me to her home many times, making me feel like I was back home. Chatting with her while cooking is the most memorable thing at UIUC.

The completion of the dissertation could not have been accomplished without the support of the committee members: Dr. Tania Ionin, Dr. Aaron Johnson, and Dr. Ryan Shosted. Dr. Ionin is also the supervisor of my qualifying paper. I benefited a lot from her careful writing style and well-organized lectures. She is my role model for being a teacher. I would like to thank Dr. Johnson for his clever questions he brought up in the discussion. His comments pushed me to think

more deeply about the issues and to vision the future research. I would like to offer my special thanks to Dr. Shosted, who gave me the opportunities to learn MRI, MATLAB, and R. Working with him in the MRI project was a great experience. I am also grateful to Dr. Sa Shen and Dr. Wood Simon for assisting me with the GAM analyses.

This work was made possible with the Title VI Grant under the PIs of Dr. Torrey Loucks and Dr. Chilin Shih, and with Cognitive Science/Artificial Intelligence Award from Beckman Institute at UIUC. My work cannot be made possible without the continuous support and the solid linguistics training in the Department of Linguistics. I appreciate that I had the chance to work in the Second Language Acquisition and Bilingualism Lab for 2 years under the supervision of Professor Silvina Montrul. I would like to acknowledge the funding support from Fulbright Scholar Program for the first two years of my PhD study. I also thank the School of Literatures, Cultures & Linguistics for the SLCL Dissertation Completion Fellowship.

I am particularly grateful for the assistance and friendship given by Di Wu and Andrew Hinderliter. Your kindness to me is beyond what words can express. I cannot imagine how boring the life in Champaign would be without you. I would like to extend my thanks to the colleagues: Hsin-yi Dora Lu, Chen-huei Wu, Shawn Chang, Meng Liu, Yinghua Yang, and Anthony C. Hegg, for their comments and encouragement throughout my PhD study. I wish to thank my Taiwanese friends: Hui-shan Huang, Kate Hsu, Wei-Fen Chen, Christine Tseng, Judy Hsu, Jimmy Chu, Chen-Hsuan Lin, Melanie Hsu, Shu-Han Chao, Frank Yung-Tin Pan, Kuan-yu Tseng, Evance Ho, Kuan Yu Ko, Jon Wang, Chang-Tse Hsieh, and Ning Hsu, for the happy gathering and their support. Assistance provided by Agatha Kuczynska, Breanne Bockwoldt, and Paula Acuna in the data collection is greatly appreciated.

My deepest gratitude goes to my parents Kuang-Ting Ning and Xiu-Mei Lee and my brother Mark. The countless encouragement you gave me during my PhD life will never be forgotten.

Finally, to my loving and supportive fiancé, Terry, your company in the past year enabled me to focus on research and is much appreciated. My heartfelt thanks go to you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

This dissertation explores the process of tone learning in Mandarin. Second language (L2) learning can be viewed as a process of transfer, where there are good reasons to hypothesize that L2 learners' sound inventory belongs to the "interlanguage" that incorporates characteristics of both L1 (their first language) and L2 (Selinker, 1972). Tone learning is unique in second language acquisition. Unlike consonant/vowel acquisition where a second language learner can borrow his first language (L1) phoneme inventory, tone learning for non-tone speakers involves construction of a brand new sound category and internal model (motor commands for the new sounds) that do not exist in a learner's L1 sound inventory. The uniqueness of tone learning has been attested in tone perception, where lexical tone perception is processed by the left hemisphere in native speakers of Mandarin but by the right hemisphere in non-tone speakers (Yue Wang, Sereno, Jongman, & Hirsch, 2003). Hence, the properties of L1 to L2 transfer and the concept of interlanguage do not apply directly to second language tone learning.

This dissertation investigates the uniqueness of internal models for tone by perturbing auditory feedback and examines how native speakers of Mandarin and L2 learners of Mandarin control their voice F0 in the presence of pitch perturbation. In what follows, I will introduce two distinct concepts relevant to the dissertation, namely, lexical tone and auditory feedback. I will also explain why audio-vocal interaction provides a powerful means to study tone learning. Following these two important concepts, I will introduce the external factors affecting tone learning that will be tested by manipulating auditory feedback.

1.1.1 Lexical Tone

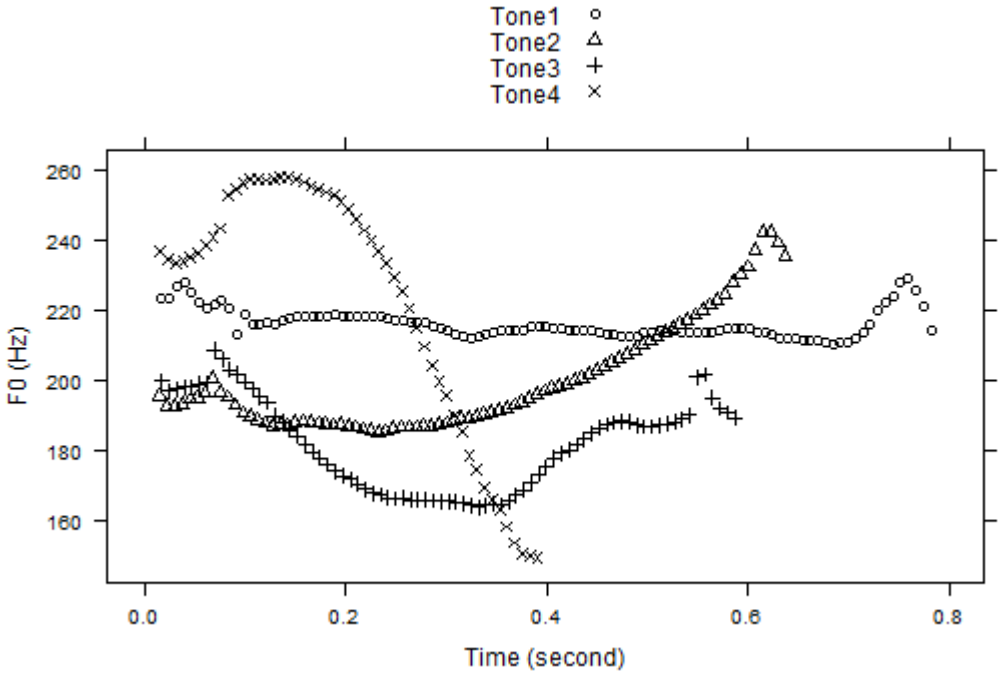
1.1.1.1 What is Lexical Tone?

Approximately seventy percent of the languages in the world may be tonal (Yip, 2002). In tonal languages, pitch is used to differentiate lexical meanings. Generally, there are two different tone systems (Maddieson, 1978; Pike, 1948). In a register tone system, commonly seen in Bantu languages, tones are differentiated by their pitch level (such as high, mid, or low pitch) (Goldsmith, 1994; Odden, 1995). In a contour tone system, typical of languages of the Southeast Asia, tones are differentiated by their distinct shape (such as falling or rising pitch) (Chao, 1933; Pike, 1948; Yip, 2002). However, there are also tonal languages that embody both register and contour tones, such as Cantonese and Wu. They are analyzed as the register tone system in Yip (1980): H and L, where each one contains level tone, rising tone, and falling tone. This dissertation focuses on Mandarin tone, which is characterized as a contour tone system that uses different pitch contours to differentiate meanings.

1.1.1.2 Mandarin as a Tone Language

Mandarin, a widely spoken tonal language, is a typical case of the contour tone system. Tones in Mandarin apply independently to each syllable and the pitch contours are comparatively complex. Mandarin syllable has a duration of approximately 180~186 ms for CV structures (Shih & Ao, 1997; Xu, 1997, 1999). The pitch contours involve four types: high level tone (tone 1), rising tone (tone 2), fall-rising tone (tone 3), and falling tone (tone 4) (Figure 1.1). For instance, in Mandarin, while *hua**4** xue**2*** means ‘chemistry’, *hua**2** xue**3*** indicates ‘skiing’; while *fang**4** huo**3*** means ‘to set on fire’, *fang**2** huo**3*** indicates ‘to protect against fire.’ Changing the pitch of a syllable alters the meaning of the word.

Figure 1.1.The F0 contours of four Mandarin tones in /ma/



1.1.1.3 Tone Learning Challenges

English, as a non-tonal language, uses pitch in a different way to express linguistic and paralinguistic information: stress in the word level (which could be multisyllabic) and intonation in the sentence level (Bolinger, 1978; Pierrehumbert, 1980). Pitch movement is typically slow considering that not every word carries stress in speech production (Hirschberg, 1988), and not every stress is realized in fast pitch movement (Pierrehumbert, 1980). However, Mandarin uses pitch at the syllable level, which has a shorter duration than a typical English word or phrase. Accurate pronunciation of the lexical tones is not only determined by the contour shape but also achieved by fast adjustment of F0. Therefore, a potential difficulty for non-tone English speakers to learn a lexical tone language like Mandarin is to have fine pitch control within a smaller domain in speech (i.e., syllable rather than word or sentence). L2 learners of Mandarin may initially rely on some hand-holding notations (mnemonic devices) to produce Mandarin tones.

However, the hand-holding notations may not be of use in producing fluent speech.

Another potential difficulty comes from hemispheric differences on pitch processing tuned by language experience. For non-tone speakers, pitch, like music, is a melody and is processed by the right hemisphere, while the left hemisphere dominates language processing (including phonemes, lexicon, syntax, and semantics). In contrast, for tone speakers, pitch is linguistically meaningful which leads to lexical contrast. Thus, pitch and other linguistic aspects, including phonemes, lexicon, syntax, and semantics, are processed by the left hemisphere for tone speakers (Van Lancker & Fromkin, 1973; Y. Wang, Jongman, & Sereno, 2001; Yue Wang et al., 2003; Wong & Perrachione, 2007).

Learning tone as a second language places unique demands on cognitive, linguistic and 'sensorimotor' mechanisms because new grammar, representations and sensations must be expressed in unique movements. Second language learners of Mandarin may have to rely on a functional brain organization shift for successful tone learning. Language learning, therefore, opens windows for observing changes in behavior and brain functions that impact the learner in multiple ways.

1.1.2 Auditory Feedback (as a Means to Study Tone Learning)

This dissertation uses auditory feedback to explore the construction of internal models for tone in second language learners of Mandarin. One way to probe the learning process is to see how L2 learners respond to perturbation in auditory feedback. This method belongs to a general category of perturbation paradigms that measures a subject's response to altered sensory environments (Abbs, Gracco, & Cole, 1984; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984; D. N. Lee & Aronson, 1974). One of the clearest examples of auditory influences on human vocalization is the Lombard effect (Lane & Tranel, 1971), which is the tendency for speakers to

increase their voice volume when speaking in a noisy environment. This change enhances the audibility of their voice but would not occur unless auditory feedback is actively monitored by the speakers. This reveals that sensory-to-motor (or sensorimotor) relations influence the regulation of a skilled and functional behavior (in this case vocalization).

Auditory feedback is important for speech and language learning. Newly learned speech sounds are stored internally as auditory targets. Auditory feedback is used to guide speech motion. After an amount of vocal practice, learners can produce acquired sounds by using feedforward commands without heavily relying on auditory feedback to guide the speech motion. In other words, unexpected movements will be modified, without actual delays of carrying out an unsuccessful movement and receiving sensory feedback about the movement (Guenther, Ghosh, & Tourville, 2006). Thus, an advanced and fluent L2 learner who has well-established internal models for speech gestures is able to correct speech movements via internal feedback loop before any mistake happens. These internal models are conceived as malleable neural systems that predict the relationships between past, current and future states of the nervous system to provide movements that are task appropriate and simulate the relationships between motor commands, motor trajectories and sensory feedback (Callan, Jones, Callan, & Akahane-Yamada, 2004; Guenther, 1995, 2006; Guenther et al., 2006; Guenther, Hampson, & Johnson, 1998; Hickok, Houde, & Rong, 2011; Jordan & Rumelhart, 1992; Kawato, 1999; Lalazar & Vaadia, 2008). The Lombard effect highlights the flexible relationships between sensory input and speech motor commands that are now thought to be instantiated by internal models in the brain.

Perturbation in auditory feedback becomes a useful tool to investigate internal models of pitch, as the response to unexpected pitch distortion reflects how internal feedback loop works to modify or avoid errors. Like singers who are trained to sing on key, tone speakers can speak on pitch. Their internalized pitch representations and strategies to respond to pitch perturbation can

be different from those in non-singers and non-tone speakers. Examining internalized pitch representations in the brain by using pitch perturbation in auditory feedback enables us to predict and evaluate learners' capability for learning Mandarin tones.

In recent years, there has been a growing interest in the field of Second Language Acquisition (SLA) regarding how suprasegmental features such as lexical tones are perceived by nontonal language speakers (Shih & Lu, 2010; Shih, Lu, Sun, Huang, & Packard, 2010; Wong & Perrachione, 2007). Apparently, tone proficiency is particularly challenging for L2 adult learners. Even for first language (L1) young children, it can take 2-10 years to achieve adult-like tone competency (2-4 years for production (C. N. Li & Thompson, 1977; Tse, 1978) and 4-10 years for perception (Ching, 1984; Ciocca & Lui, 2003; Connie Suk-Han Ho & Bryant, 1997; Lin, Wang, & Shu, 2012)). Research on adults L2 tone learning has focused mainly on tone perception, but tone production proficiency (see (Shih & Lu, 2010)) and the role of auditory feedback in L2 tone acquisition are not understood. Empirically testing the effect of auditory feedback on speakers learning tonal language is a crucial step toward better understanding of the connection between tone perception and production. It is well-established that the control of suprasegmental features relies on hearing one's own speech (auditory feedback). For instance, people with postlingual deafness have difficulty regulating loudness and intonation, while their production of phonemes remains intact (Binnie, Daniloff, & Buckingham, 1982; Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991; Waldstein, 1990). Thus, learning and monitoring lexical pitch, which is a suprasegmental feature, has to rely on auditory feedback. Second, by looking at relations between articulatory commands, acoustic characteristics and their trajectories and sensory feedback, we can investigate whether the underlying representations for vocal control in speakers are reshaped by language learning experiences.

My dissertation builds connections between SLA research and speech physiology research

to address a central question in both disciplines: how does effective language learning reshape the underlying language system. An innovative paradigm that imposes unexpected pitch-shifts (pitch-shift paradigm) indicates that adult speakers produce pitch-shift corrections to unexpected auditory feedback shifts (Burnett, Freeland, & Larson, 1998; Burnett & Larson, 2002). The upward or downward compensatory corrections in F0 are used to correct the pitch errors that speakers hear in their voice feedback. It provides compelling evidence for an error monitoring system in the brain to regulate F0. Recent pitch-shift studies have recently found unique pitch-shift response patterns in native speakers of Mandarin, which have opened many possibilities to Mandarin tone research and applications (Burnett et al., 1998; Hain et al., 2000; Jones & Munhall, 2000, 2002; Larson, Burnett, Bauer, Kiran, & Hain, 2001; Xu, Larson, Bauer, & Hain, 2004). Native Mandarin speakers showed compensatory pitch-shift responses with shorter latencies and larger amplitudes compared to native English speakers. This unique pattern of pitch-shift responses indicates native Mandarin speakers regulate their voice F0 during tone production by a fast adjustment mediated by internalized pitch representations. Pitch-shift responses, therefore, could serve as a probe for examining the attainment of native-like tonal representations in the human brain.

1.1.3 External Factors for Tone Learning

Apart from considerations directly related to the 'auditory-to-motor' loop, there are other external factors that could influence tone/pitch learning, including language experience, musicianship, and non-auditory sensory feedback. Speakers' language experience (Halle, Chang, & Best, 2004; Xu, Gandour, & Francis, 2006) and more recently identified musical experience (e.g., trained singers (Cooper & Wang, 2012; C.-Y. Lee & Hung, 2008)) seem to provide an advantage in behavioral tone identification that is reflected in different neurological responses to

linguistically relevant pitch contours (Chandrasekaran, Krishnan, & Gandour, 2007b; Krishnan, Xu, Gandour, & Cariani, 2005). Speaking a tonal language may help one to learn another tonal language, though the extent of facilitative effect depends on the complexity of one's native and target tone systems. Having musical training could help one to learn a tonal language as well. The musicians' abilities in pitch decoding can be transferred across domains. In other words, the pitch experience obtained in one domain (music) may facilitate the learning capability in another domain (language) (Bidelman, Gandour, & Krishnan, 2011).

Additionally, traditional verbal feedback from teachers' description of students' production errors and examples of tone production from teachers in classrooms may have limitations. The successful language learning not only depends on instructors' ability to communicate but also on students' ability to interpret the instructors' words. It has been shown in singing that visual cues in pitch contour simultaneously presented with auditory stimuli enhances pitch accuracy (Howard et al., 2007; Kawase et al., 2009; Thorpe, 2002; Wilson, Lee, Callaghan, & Thorpe, 2008). Tracing back to the late 19th century, Alexander Graham Bell had already experimented with speech visualization to teach deaf people to speak (J. Flanagan, 1972). The phonograph that draws vibrations from human voice helped deaf students visualize the correct "shape" of the sound they were trying to make so that they could compare their own speech with it. Visual feedback for teaching suprasegmentals (tone and intonation) has also been shown beneficial to foreign language students (Burke, 1996; Chun, 1989; J. Flanagan, 1972; Weltens & Bot, 1984). Thus, provision of specific real-time visual feedback that prompts suprasegmental voice changes could help L2 learners to stabilize their pitch. Examining cross-modal sensory feedback in this dissertation can lead to real-world tools for language instruction and complement traditional classroom instruction.

1.2 Research Questions

This dissertation investigates external factors and internal factors for Mandarin tone perception and production in four populations, including naïve (no exposure to tonal languages), L2 adult learners, trained vocalists, and native speakers of Mandarin. The first study of the pitch-shift response and non-/linguistic tone discrimination investigated whether pitch-shift responses were related to tone discrimination ability. The second study examined the pitch-shift responses to Mandarin tone sequences and employed an adaptive test to understand how Mandarin tone discrimination is associated with Mandarin experience, musicianship, and pitch-shift responses. The third study tested the effect of real-time visual feedback on suprasegmental tone control.

The following research questions and hypotheses are tested:

1) Perception-Production Relationship:
    a) Is Mandarin tone discrimination ability related to basic nonlinguistic tone discrimination ability in naïve speakers, L2 learners, musicians, and native speakers? (Study 1-2)
        ➔ ***Hypothesis 1***: L2 learners and musicians will display response patterns that resemble Mandarin speakers. Discrimination ability of the L2 learners, musicians and native speakers will be superior to that of naïve speakers without musical experience.
    b) Are pitch-shift responses to the simple vowel /a/ or to Mandarin disyllabic tone sequences related to Mandarin tone discrimination ability in naïve speakers, L2 learners, musicians, and native speakers? (Study 1-2)
        ➔ ***Hypothesis 2***: Audio-vocal response amplitude and latency (response onset

time) will be significantly correlated with Mandarin tone discrimination ability in all four populations, with characteristic patterns that can be used in predictive models.

- ➔ ***Hypothesis 3***: Pitch-shift responses vary between Mandarin tones and the responses to Mandarin tones differ from the responses to the simple vowel.

2) External-Internal Relationship:

a) Does musical experience promote Mandarin tone discrimination and pitch-shift responses that resemble experience with a tonal language? (Study 2)

- ➔ ***Hypothesis 4***: Audio-vocal response amplitude and latency (response onset time) will be correlated to musical experience.

b) Are pitch-shift responses facilitated by real-time visual feedback of vocal fundamental frequency? (Study 3)

- ➔ ***Hypothesis 5***: Real-time visual feedback will decrease the amplitude of pitch-shift responses in native speakers, L2 learners, and musicians, but not in naïve speakers.

1.3 Outline

Chapter 2 reviews the literature on tone learning, speech production models and the pitch-shift paradigm followed by a summary of research on pitch processing and pitch learning. Chapters 3-5 present the results of Study 1-3, respectively. Study 1 explored the relationship between pitch-shift responses and non-/linguistic tone discrimination ability. Three tasks were implemented in Study 1: a nonlinguistic tone discrimination task, a Mandarin tone discrimination task, and a pitch-shift task where participants were instructed to produce the simple vowel /a/. Study 2 investigated the stimulus specificity of pitch-shift responses and their relation to

non-/linguistic tone discrimination ability. Three tasks were conducted in Study 2: a nonlinguistic tone discrimination task, an adaptive Mandarin tone discrimination task, and a pitch-shift task where participants were instructed to produce Mandarin disyllabic tones (T11, T12, and T14) and the simple vowel /a/. Study 3 examined the role of real-time visual feedback in suppressing pitch-shift responses and stabilizing voice F0. Two tasks were administered in Study 3: a nonlinguistic tone discrimination task and a pitch-shift task where participants were instructed to produce Mandarin high level tone (Tone 1) and the simple vowel /a/. Chapter 6 provides a discussion on tone discrimination ability and audio-vocal responses. Chapter 7 concludes the dissertation.

CHAPTER 2
LITERATURE REVIEW

This chapter reviews young children's and adults' tone acquisition and their performance on tone perception (Section 2.1), followed by a discussion of relevant speech production models and internal models that account for feedback influences on motor control (that are central to speech learning) (Section 2.2). The pitch-shift paradigm which imposes an unexpected pitch perturbation in auditory feedback will be reviewed in Section 2.3. Section 2.4 discusses some external factors such as linguistic experience, musical background, and real-time visual feedback that may affect pitch processing.

2.1 Tone Learning

2.1.1 First Language (L1) Acquisition

Li and Thompson (1977) reported the first systematic study of tone acquisition in which they examined the tone acquisition process in Mandarin-speaking children. Their cross-sectional (between age 1;6 and age 3;0) and longitudinal (7 months) study on 10 Mandarin-speaking children reveals that children produced a phrase or a word with the appropriate tone pattern which was segmentally different from the adult form. This suggests that the tone system is acquired earlier than the segmental system. The same pattern is apparent in Cantonese children (Tse, 1978), based on the author's longitudinal record (from age 2;8 to age 5;3) of his son Y.L.. Y.L. started to acquire tones from his one-word stage and the acquisition of tones was complete at the two- or three-word stage, which is earlier than the completion of segmental acquisition. As for the acquisition order of the four lexical tones, for example in Mandarin, the high (T1) and falling (T4) tones are acquired earlier and more easily than the rising (T2) and dipping (T3) tones

(C. N. Li & Thompson, 1977). Adopting the arguments in Ohala & Ewan (1973), they claimed that rising tones are acquired later than falling tones since i) raising pitch requires greater physiological effort than lowering pitch, and ii) rising tones are perceptually more difficult to detect than falling tones.

While Li and Thompson (1997) and Tse (1978) argue that the tone acquisition in Mandarin-speaking or Cantonese-speaking children could be completed by two- or three-word stage, Ching (1984) and Ciocca and Lui (2003) argue that children's performance reached an adult level only at 10 years of age. Their experimental findings of tone perception on (4, 6 and 10 years old) Cantonese-speaking children and adults show that the tonal system is not fully acquired by young children even though they use lexical tones quite early. Children's performance improved considerably from age 4;0 to age 6;0, and from age 6;0 to age 10;0. By 10 years of age, children performed as accurately as adults. The achievement of tone perception may have to do with metalinguistic awareness of onset and rhyme, which is developed later at school age. Chinese-speaking children are able to detect tone independently of onset or rhyme by age 5 (Connie Suk-Han Ho & Bryant, 1997) or by first grade (Lin et al., 2012; Siok & Fletcher, 2001). Rhyme plays an important role in facilitating lexical tone processing, as Chinese-speaking children performed significantly better in the condition where the syllables had the same rhyme but different onset than in the condition where the syllables had the same onset but different rhymes (Lin et al., 2012).

Tone awareness is crucial in learning words in tonal languages such as Chinese. Weakness in tone awareness may give rise to a developmental delay in word reading and writing. Chinese children (particularly Cantonese-speaking children) with developmental dyslexia have a special difficulty in tone processing, which is potentially more severe than their difficulty in processing onsets and rhymes (Connie Suk-Han Ho & Chan, 2002; Coonnie Suk-Han Ho, Chan, Lee, Tsang,

& Luan, 2004; W.-S. Li & Ho, 2011). In Li and Ho (2011), for instance, children who scored 1.5 or more standard deviations below the mean of same-age normal readers were regarded as having a deficit in tone perception. Their results show that 35% of the dyslexic children had difficulty in discriminating tone of a known word whereas only 20% of the dyslexic children showed difficulty in discriminating rhymes.

2.1.2 Second Language (L2) Acquisition

Critical period hypothesis claims that acquiring a language would become more difficult and effortful after a certain age (Lenneberg, 1967; Penfield, 1959). The cerebral lateralization and the loss of neurological plasticity of the brain after puberty may reduce the ability to learn foreign language (Lenneberg, 1967). Various models have been postulated to account for the different patterns of results in adult foreign speech perception, including Best's Perceptual Assimilation Model (Best, McRoberts, & Goodell, 2001), Flege's Speech Learning Model (Flege, 1995), and Kuhl's Native Language Magnet Model (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992).

Best's Perceptual Assimilation Model argues that the influence of native L1 categories on L2 segmental processing could be inhibitory or facilitative, depending on whether non-native contrasts a listener perceives fit into a single native L1 phonological category. Discrimination of a non-native contrast could be near perfect if the contrast also occurs in the learner's native language (e.g., English-speaking adults could differentiate Zulu's voiceless versus voiced lateral fricatives /ɬ/-/ɮ/ as voicing contrast in fricatives occurs in English as well). Discrimination of a non-native contrast could be lower but still good if the contrast can be interpreted as good versus poor exemplars of a single L1 native phoneme (e.g., Zulu's voiceless aspirated versus ejective velar stops /kʰ/-/k'/ perceived by English-speaking adults). However, discrimination of a

non-native contrast could be very difficult if the contrast matches equivalently to a single native phoneme (e.g., Zulu's plosive versus implosive voiced bilabial stops /b/-/ɓ/ perceived by English-speaking adults) (Best et al., 2001).

Flege's Speech Learning Model argues that the phonology in one's native language may filter out phonetically important (acoustic or gestural) features of non-native contrasts so that second language learners fail to discriminate the contrast in a foreign language. Production of L2 sounds would be incorrect if no accurate perceptual targets guide the learning of L2 sounds (Flege, 1995).

Kuhl's Native Language Magnet Model argues that perceptual magnets, which was assessed by the degree to which 6-month old infants did not detect a difference between a prototype and its variants, would be strong only for native-language phonetic prototypes. For instance, American 6-month old infants perceived the American English /i/ variants as identical to its prototype /i/ on 66.9% of the trials while perceived the Swedish /y/ variants as identical to its prototype /y/ only on 50.6% of the trials. The native language prototype's magnet effect explains why adults fail to discriminate two speech sounds from a foreign language when both sounds fit into a single prototype in their native language (Kuhl et al., 1992).

In general, these models do not focus on how non-native suprasegmentals would be perceived, nor do they focus on how and whether nonnative sounds could be learned. Recent research on perceptual reorganization for tones in infants has shown that there is a decline in (Thai) lexical tone discrimination between 6 and 9 months of age for English-learning infants (Mattock & Burnham, 2006) and French-learning infants (Mattock, Molnar, Polka, & Burnham, 2008), but no such decline for Chinese-learning infants (Mattock & Burnham, 2006). The discrimination of nonnative tones deteriorates between 6 and 9 months rather than between 4 and 6 months (Mattock & Burnham, 2006; Mattock et al., 2008). The findings suggest that lexical

tone perception is shaped by exposure to a tonal language and that the development of lexical tone begins in early infancy starting around 6 months of age.

As for adults' tone perception, Hallé, Chang, and Best (2004), as an initial proposal, examined whether French listeners perceive tonal differences in a linguistic way similar to Mandarin listeners. Using the AXB identification test (A and B corresponding to two endpoints in two possible orders and X varying from one endpoint to the other along the eight steps of the continuum), French listeners had more trouble than Mandarin listeners in categorizing unambiguous tones represented by the endpoints. French listeners had lower accuracy and longer reaction times on the endpoint trials. However, for ambiguous tones, French listeners were better at discriminating whether an intermediate tone is closer to one endpoint than to the other. In addition, in the two-step AXB discrimination task (either X=A or X=B), Mandarin listeners' sensitivity to tonal differences were biased by tonal category, where the categorical boundary was the sharpest for tone 1–2 pair, sharp for tone 2–4 pair and the fuzziest for tone 3–4 pair. In contrast, French listeners had no such bias. The results indicate that French listeners show a sensitivity to tone contour variations but the judgments are determined by psychophysical factors rather than by linguistic factors. Mandarin listeners' perception is affected by language experience. The language experience effect also extends to the categorical perception of homologous nonspeech harmonic tones (Xu, Gandour, et al., 2006).

Auditory training has been used to assess the trainees' improvements and learnability on the acquisition of non-native suprasegmental contrasts. English-speaking adults were able to identify four Mandarin lexical tones in non-lexical contexts (words presented in isolation) after eight sessions (40 minutes each) of training, in which the identification accuracy increased by an average of 21% and the improvement was maintained for 6 months after training (Yue Wang, Spence, Jongman, & Sereno, 1999). Wong and Perrachione (2007) as a complementary study to

Wang *et al*. (1999) trained English-speaking amateur musicians and nonmusicians to identify Mandarin tones in lexical context. Trainees learned to associate the image with 1 of 18 pseudowords during the training (30 minutes each session). Training was terminated and regarded as successful learning when participants showed at least 95 % accuracy for two consecutive sessions, but terminated and regarded as less successful when participants failed to improve by at least 5% accuracy for four consecutive sessions. Learning success in identifying pitch pattern in words was found to be predicted by the ability to identify pitch pattern in a non-lexical context (pre-training tone identification result) and by musical experience. It suggests that basic auditory discrimination ability, possibly shaped by musicianship, is related to the success in tone learning.

Training method may have an effect on tone learning. Shih *et al*. (2010) evaluated three tone training programs on L2 learners of Mandarin using a large set of training materials. The materials included 15 monosyllabic Mandarin words with 4-way tonal contrasts. Each word was recorded at 11 talker-to-listener distances (TLD, see Cheyne, Kalgaonkar, Clements, and Zurek (2009); Pelegrin-Garcia, Smits, Brunskog, and Jeong (2011)) by three native speakers and repeated 4 times, which yielded 7920 tokens. The distance introduced speech variation (exaggerated, normal, and reduced tokens) and represented the difficulty level of the materials in the data bank. The first group received clearly articulated and acoustically less varied speech sounds that were recorded at distance 5 and 6. The second group received all tokens from the data bank but in a random order. The third group received all tokens through an adaptive training program where the difficulty of the stimulus was chosen based on a learner's performance. There is a control group where learners did not receive tone training. The tone identification results from 36 Mandarin learners showed a non-linear response to speech recorded at different TLD. Speech recorded at a close distance was difficult for learners due to reduction, while those

recorded at a long distance was harder due to exaggeration. Speech files recorded with a TLD distance of 8 feet (step 4 of Shih's et. al.) have the highest identification rate by learners. Learner's tone identification accuracy decreased linearly from step 4 (TLD of 8 feet) to step 11 (TLD of 20 feet), if an aberrant bump at step 10 was excluded. Only the adaptive training program facilitated the learning of tone variants in the brief two and a half hour training program.

The training studies point to the possibility that adults' perceptual mechanisms have more plasticity than was previously suggested by infants' research. Although perceptual reorganization may be achieved by training, it remains unclear whether the non-tonal language speakers actually reshaped tonal representations in the brain and whether their tone production was altered to resemble native speakers. In order to bridge the gap between perception and production, this dissertation will focus on the relationship between tone discrimination ability and audio-vocal responses to pitch-shift stimuli. In the next subsection (2.2), I will review some schools of thought in modeling speech production, which may provide insight for examining second language production.

2.2 Speech Production

Speech production models have been proposed to account for the speaker's selection of a word, the corresponding articulatory gestures, and spontaneous or induced speech errors. Notably, Levelt's lexical access model, Dell's connectionist model, and Browman and Goldstein's gestural patterning model all focus on how speakers generate spoken words (section 2.2.1). These models propose the relevance of feedforward planning mechanism to word production. However, none of the word production models captures the role of the peripheral feedback system in shaping speech motor control. Feedback from periphery used for correcting movements to reach desired

goals is also essential, particularly when learning a language. The importance of periphery feedback mechanism is proposed in internal models by Guenther (1995, 2006) This dissertation will rely considerably on internal model and its variants which involve feedback loops to account for the accuracy of speech movements (section 2.2.2).

2.2.1 Word Production Models

The lexical access model in Levelt, Roelofs, and Meyer (1999) proposes that the production of a word is conceived as a staged process, from conceptual preparation, lexical selection, morphological and phonological encoding, phonetic encoding, to the initiation of articulation. The self-monitoring mechanism monitors the internal speech, which affects (phonological or phonetic) encoding duration. Compared to the feedback system in internal models which will be reviewed in section 2.2.2, the self-monitoring mechanism in Levelt's model does not capture short-latency influences on feedback on spoken output. The lexical selection mechanism in the theory developed by Roelofs (1992) attaches a lemma stratum to the conceptual stratum. The lemma node receives the activation spread from the conceptual node and then sends the activation downward to the form level which connects to phonemes.

Unlike the lexical access model which has a discrete and unidirectional network, Dell's (1986) connectionist model is a bi-directional and interactive system, where activation spreads both ways. The connections in the connectionist model are weighted, as each node can collect and sharpen activation from different sources. An example for this weighting is the lexical bias happening in speech errors. The lemma node *cat*, for instance, is activated and spreads the activation to its phoneme nodes /k/, /æ/, and /t/. The semantic features of *cat*, such as animal and mammalian, co-activate the lemma node *dog* or *rat*. The phonemic features along with the semantic features of *cat* boost a better chance for *rat* to emerge as a speech error than for

semantically related *dog*, phonologically related *mat*, or even for nonword *gat*. Feedback can potentially occur between levels but the type of feedback is not specified and is not sensorimotor in nature.

Browman and Goldstein's (1989) gestural patterning model argue that speech is defined in terms of movement patterns regulated to achieve speech goals. The movement patterns are a product of the interaction between articulator variables and tract variables. The tract variables are associated with context-independent gestures. The articulator variables, however, are associated with context-dependent gestures. Thus, for a given vocal tract configuration, there can be various articulatory involvements. In the gestural patterning model, activation of gestural units at the intergestural level can account for the relative timing for gestural movements. For example, the initial /p/ requires a bilabial closure and a glottal opening for voiceless speech. In fluent speech, overlap in gestures can be observed, which accounts for sound deletion (e.g., *perfect memory* /pəˈrfəkt 'mɛməˈri/ → [pəˈrfək'mɛməˈri]) or sound assimilation (e.g., *seven plus* /sɛvn̩ 'plʌs/ → [sɛvm̩ 'plʌs]) (Browman & Goldstein, 1989). A role for feedback is not specified in this model.

The lexical access model and the connectionist model have succeeded in modeling many aspects of speech production behaviors, including semantic, orthographical and phonological influences on lexical selection, as well as sequencing errors. However, some crucial components are missing regarding how speakers control the speech system. The gestural patterning model has succeeded in modeling biomechanical movements and may be applied to articulatory synthesis system. However, it has a serial order problem in the sense that it does not explain the changes that are affected by non-neighboring segments. None of the three models incorporated sensory feedback. Feedback from the periphery provides a way to correct for unexpected changes in the oral tract and thus plays an important role in speech motor learning. It has been established from studies of pre-linguistically deaf children that learning an 'aural/oral' language depends heavily

on auditory feedback to adjust production and to build internal representations. In what follows, I will review internal models that involve a feedback system and discuss the importance of auditory feedback to speech production.

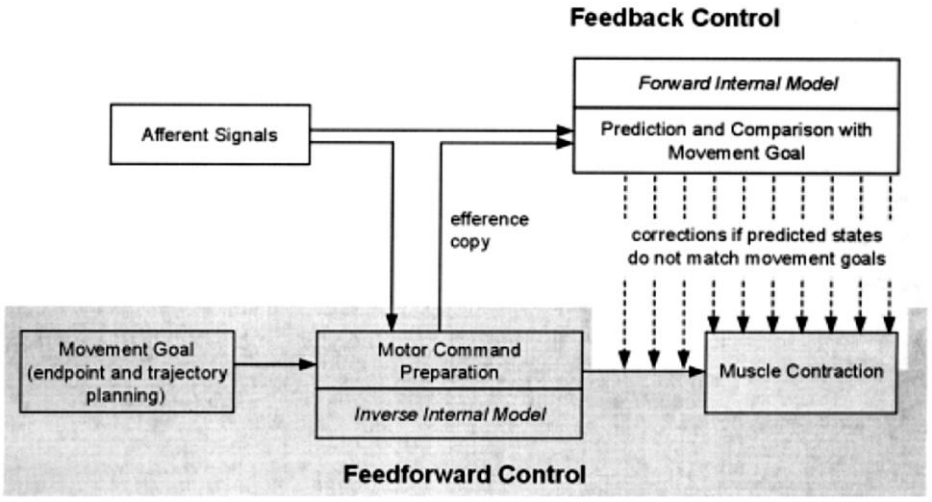
2.2.2 Internal Models

Articulatory movements for linguistic gestures, including Mandarin tone gestures, are rapid and are typically completed within a syllable, i.e., within 200-300 milliseconds. However, there are critical delays in the generation of movements in the central nervous system. Feedback is too slow to alter ongoing speech. This rationale had led to the notion of strictly feedforward control for speech production where feedback was irrelevant. Yet more recent experiments show that sensory feedback can alter ongoing speech production at the level of formant production (Cai, Ghosh, Guenther, & Perkell, 2010). With new evidence from these studies, sensory feedback is now seen as an important component providing oral tract context information that shapes the planning of speech production. By learning from sensory input, a mature speech production system can anticipate the delays in feedback to shape upcoming speech commands. This reconciliation of speech production with feedback and the associated delays has inspired innovation in speech production theories. One noted theoretical advance has been the proposal that internal models within the nervous system can simulate the relation between articulatory commands, acoustic trajectories and sensory feedback (Guenther et al., 2006; Lalazar & Vaadia, 2008; Perkell et al., 1997). This section defines and reviews the concept of internal models, and then proposes that the Directions into Velocities of Articulators (DIVA) model of Guenther (2006) provides a rational for the design and interpretation of the proposed studies. Evidence for the role of auditory feedback and existence of feedforward control as proposed by the DIVA model is also reviewed.

2.2.2.1 The Internal Model Concept

Internal models (see general diagram in Figure 2.1) are neural mechanisms that actually simulate the association between motor commands, resulting motor trajectories and associated sensory feedback (Jordan & Rumelhart, 1992; Kawato, 1999; Lalazar & Vaadia, 2008). It basically means the brain develops an 'internalized' representation of a motor goal. By modeling the relationships between the movement execution and outcome, the brain can predict the muscle forces required for the task, simulate the expected sensory feedback, and make feedback based corrections without the long delays of the actual feedback. Therefore, the roles of motor commands and sensory correction can be reconciled.

Figure 2.1. Internal model



(Republished with permission of AMERICAN SPEECH - LANGUAGE - HEARING ASSOCIATION from Max, Guenther, Gracco, Ghosh, and Wallace (2004); permission conveyed through Copyright Clearance Center, Inc.)

It has been predicted that several general types of internal models function together to produce motor commands. *An Inverse Internal Model* takes the desired movement goal and

prepares the motor commands required to accomplish it. An efference copy generated by the motor control system is the outcome of an inverse internal model which contains the motor commands for the task. *The Forward Internal Model* predicts the sensory state resulting from motor commands (afferent signals), which is essentially the predicted outcome of a movement. A comparator then checks if the predicted outcome of the movement corresponds to the motor commands. If so, the motor commands are issued to the muscles. If there is a mismatch, then an error signal is generated to modify the motor commands. Because the mismatch is only represented internally, modifications to the motor commands are made without requiring the actual feedback from the movement.

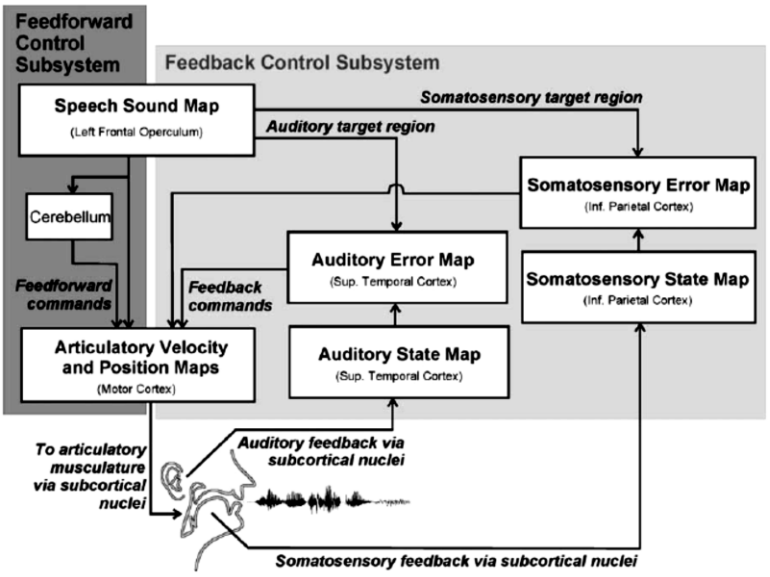
Internal models typify the highly skilled movements of athletes or musicians in which rapid, complex sequences of movements are successfully executed without the apparent need to monitor these movements. The internal model hypothesis has been used to account successfully for coordination between reaching and grasping (J. R. Flanagan & Wing, 1997), hand-eye coordination (Held & Hein, 1958), and speech production (Perkell et al., 1997). Infants going through the process of learning to speak reveal that the motor apparatus of the jaw, tongue and vocal folds can be specifically and rapidly shaped for this very sophisticated function. These researchers suggest that motor learning relies on forward models to give instantaneous estimates of feedback and to generate sensory error signals, which in turn guide and modify motor commands in the inverse models.

Learning or building the internal model is a motor learning process. For speech, this process takes place in the period between 6 months to 7 years in which a child gradually becomes an 'expert' speaker. As the movement goal is first attempted and practiced perhaps with mistakes, the relationships between the movement and sensory feedback are encoded within the sensorimotor system. As the person becomes more successful at performing the movement, the

movement can be executed more rapidly and becomes less dependent on actual sensory feedback. Essentially, the brain has learned the expected feedback and uses internal predictions of the feedback to modify and update motor commands without the delays from actual feedback.

With specific attention to speech motion, Guenther (2006) argues that production of speech sounds requires three types of information: auditory, somatosensory, and motor, which are encoded in the temporal, parietal, and frontal lobes of the cerebral cortex, respectively. Guenther has developed a unique model that uses computer simulations to predict how motor commands result in an acoustic signal. His model, called the Directions into Velocities of Articulators (DIVA) model, that mimics neural mechanisms, sensory feedback and motor commands, is implemented in the learning and mature production of an acoustic signal (Guenther, 1995; Guenther et al., 2006; Guenther et al., 1998). DIVA mathematically implements internal models to generate motor commands in the model via three control subsystems: a feedforward control subsystem, an auditory feedback control subsystem, and a somatosensory feedback control subsystem (Figure 2.2).

Figure 2.2. The DIVA model



(Reprinted from Guenther (2006), with permission from Elsevier)

As the *feedforward control* element of the DIVA is learned, it comes to constitute motor commands required for speech production. In the training phase which is analogous to infant babbling, the model is tuned to map onto target regions of speech sounds. The use of target regions, rather than point targets, in the DIVA model provides a way to account for variability in speech, such as contextual variability, coarticulation, and speaking rate effects (Guenther, 1995). *Auditory feedback* from self-generated speech sounds signals the difference between the auditory target and the incoming auditory signal. If the current auditory feedback falls outside the target region, it would activate auditory error cells in the posterior superior temporal gyrus and planum temporale to modify the speech (Tourville et al., 2005). Similarly, *somatosensory feedback* from speech articulators indicates the match/mismatch between the target and actual somatosensory state. If there is unexpected gesture such as jaw perturbation, somatosensory error cells in the postcentral gyrus and supramarginal gyrus would be activated to correct the errors (Tourville et

al., 2005). The corrective commands issued by auditory feedback control and somatosensory feedback control are stored in the feedforward command for the next attempt. Once the DIVA model has learned appropriate motor commands for a speech sound, it can correctly produce the sound by using just the feedforward commands projected from the premotor cortex and the primary motor cortex.

The DIVA model has advantages over other models for several reasons: 1) The model simulates the babbling phase of speech; 2) the model shows evidence for spontaneous co-articulation; and 3) the neural systems for motor movement, sensory feedback and sensorimotor corrections are explicit in the model. Evidence for internal models that underlie skilled speech production has been shown with perturbation paradigms such as the formant shift (Houde & Jordan, 1998, 2002) and pitch shift paradigms (Bauer & Larson, 2003; Burnett et al., 1998; Burnett & Larson, 2002; Donath, Natke, & Kalveram, 2002; Jones & Munhall, 2000, 2002; Larson, Burnett, Kiran, & Hain, 2000). The upward or downward compensatory vocal response to the perceived errors is rapid with reflex-like properties. It is difficult to suppress responses even if a speaker is aware of the perturbation. In brain imaging and neurophysiological studies, it has been shown that neural activation (anterior cingulate cortex, auditory cortex and the insula in fMRI; P1-N1-P2 in ERP) are less suppressed in response to perturbed auditory feedback compared to non-shifted feedback, suggesting that the brain can detect the error (the perturbation) and attempt to execute a command to correct the error (Behroozmand & Larson, 2011; Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Eliades & Wang, 2008; Heinks-Maldonado, Mathalon, Gray, & Ford, 2005; Parkinson et al., 2012). Thus, the pitch-shift response provides compelling evidence for an error monitoring system in the brain responsible for the regulation of vocal pitch production.

2.2.2.2 Speaking-induced Suppression: Evidence for the Role of Auditory Feedback

Important evidence for the role of auditory feedback in speech motor control is the speaking-induced suppression (SIS) phenomena. It involves a reduced brain response in auditory cortical regions during self-produced speech compared to listening to externally-produced speech (Houde, Nagarajan, Sekihara, & Merzenich, 2002). According to the DIVA model, this reduction indicates auditory feedback from actual incoming speech signals is being compared with predicted speech outcomes derived from efference copies of motor commands. When the efference copy matches what is being produced, as in self-generated speech, auditory brain activity is reduced. If there is a mismatch that deviates from the expectation derived from the efference copy, then there is increased brain activity to signal the mismatch.

Several experimental reports have provided evidence for SIS. The M100 auditory response in magnetoencephalography (MEG) is greatly suppressed in the speaking condition where participants produced a schwa, compared to the tape playback condition where participants only heard a tape-recording of their own schwa. However, when participants' speech feedback was altered by adding white noise during speech production, the auditory suppression to self-produced speech compared to tape playback condition disappeared (Houde et al., 2002). The absence of suppression in the noise situation is consistent with the hypothesis that the speaking-induced suppression results from matching actual auditory signal with internal expected input.

Rapidity and complexity of utterances may also affect the magnitude of speaking-induced suppression. In a MEG study by Ventura, Nagarajan, and Houde (2002)**,** participants were instructed to speak the designated speech targets and then listen to the playback. The speech targets included simple speech /a/, rapid speech /a-a-a/, and more complex speech /a - a-a - a/, which were randomized in the experiment. Speaking-induced suppression percent differences

were greatest with simple speech, smaller with rapid speech, and smallest with complex speech. This finding also supports the claim that auditory feedback plays an important role in speech production. Matching incoming auditory feedback with internal expected outcome is facilitated when the utterance is simple and static than when it has dynamic articulations, as dynamic acoustic signals could generate larger prediction errors in speaking.

2.2.2.3 Sensorimotor Adaptation: Evidence for the Existence of Feedforward Control

The feedforward subsystem in the DIVA model plays an important role in storing and generating motor commands. Mature speech requires precisely timed feedforward commands that are not dependent on continuous feedback for generating fluent speech (i.e., error-free at typical and fast rates). In the DIVA model, learned speech production goes directly through the feedforward route (Guenther, 2006). Evidence for this comes from several observations. Adults with post-linguistic deafness can still produce highly intelligible speech despite the absence of auditory feedback. Somewhat similarly, normal adult speakers can speak highly intelligibly under masking noise or under altered movement conditions, using anecdotal examples such as speaking while eating or with a cigar in the mouth.

Current experimental evidence indicates that feedforward control is not static but can be updated by changes in feedback. An adaptation study used gradual but slight alterations in auditory feedback to modify speech output in adults. The formants of English vowel /ɛ/ in CVC context such as ‘pep’ were increasingly altered at the training phase until reaching maximal alteration strength (Houde & Jordan, 1998, 2002; Villacorta, Perkell, & Guenther, 2007). Participants compensated for the altered feedback and retained the compensation when the feedback was blocked by noise. The adaptation, also called after-effects, can generalize to other vowels that are not included in the training set (i.e., ‘pip’) and to contexts that involve different

consonants (i.e., ‘peg’). The adaptation shows that the corrective commands attempting to bring the production back to the expected output have modified the motor commends in the feedforward subsystem, which can be directly retrieved to produce the next attempt.

The phenomenon of adaptation and generalization in static vowels also happens to time-varying formant frequency trajectories such as the triphthong /iau/ in Mandarin (Cai et al., 2010). However, Mandarin speakers had weaker compensation (about 16%), compared to 40% in Villacorta et al. (2007) and 54% in Houde and Jordan (2002). Cai et al. (2010) argue for two possible explanations: (i) The vowel space is not as crowded in Mandarin with 7 monophthongs compared to English with 10 monophthongs. Thus, smaller auditory errors and smaller compensatory changes are elicited in Mandarin with the same amount of perturbation; (ii) time-varying articulatory trajectories may be inherently less dependent on auditory feedback than static articulation in monophthongs, as the somatosensory subsystem is available for adjusting the information and optimizing the performance.

2.2.3 Interim Summary

Section 2.2 reviews speech production models and pays specific attention to internal models that can capture the role of peripheral feedback system in speech motor control. Auditory feedback (listening to the sound) is the primary sensory modality for mastering pitch contrasts, as visual feedback of phonation is not available naturally. In Section 2.3 below, I will discuss internal factors that may affect pitch processing. By introducing disturbance in auditory feedback (the pitch-shift paradigm) we are able to examine internalized tone representations and how we control the laryngeal movement in the face of perturbation. Doing so may shed light on the possibility of reshaping internalized pitch representations in second language learners. In addition to this, Section 2.4 reviews some external factors that could affect tone learning,

including language experience, musical experience, and real-time visual feedback.

2.3 Internal Factors on Pitch Processing: Examination of Internalized Tone representation by the Pitch-shift Paradigm

The compensatory responses to formant frequency shifts discussed in Section 2.2.2.2 are also observed when pitch feedback is shifted. In the 'pitch-shift paradigm', unexpected pitch-shifts are fed back to speakers during vocalization. Essentially, the speakers hear a brief artificial change in the pitch of their vocalizations. The direction of pitch-shift stimulus (PSS) can be upwards (increase) or downwards (decrease). Speakers typically respond rapidly to the pitch-shift stimulus with either a compensatory (opposite direction to the stimulus, which is a corrective response) or a following response (same direction as the stimulus). Like formant perturbations, speakers typically produce compensatory pitch-shift responses to these unexpected shifts (Burnett et al., 1998; Burnett & Larson, 2002; S. H. Chen, Liu, Xu, & Larson, 2007; Hain, Burnett, Larson, & Kiran, 2001; Larson et al., 2000; Liu, Auger, & Larson, 2009). The correction, i.e., the compensatory response, is made because the speakers would get tricked into thinking they have made an error in the vocalization. The PSS can affect a single syllable so its influence on a tonal language could be more profound and specific compared to non-tonal languages. The following section reviews how the pitch-shift paradigm reveals audio-vocal interactions in the control of fundamental frequency production.

2.3.1 Pitch-shift Responses to Speech, Nonspeech, and Singing

Over the past 15 years, the pitch-shift paradigm has been utilized to examine different speaking and feedback conditions, including speech (word stress or phrasal/sentential intonation) and singing. The effect of the pitch shift perturbation is usually measured in terms of the

magnitude of the pitch correction (amplitude measured in cents; cents = $1200*\log_2$(F0/baseline F0)), how long it takes for the pitch correction to be initiated (latency measured in millisecond (ms)) and the temporal peak of the corrective response (latency measured in ms).

Response magnitudes during singing (singing the nonsense word /'ta:tatas/, mean 66.1 ±30.4 cents) were higher than while speaking (speaking the nonsense word /'ta:tatas/, mean 46.7 ±37.3 cents) (Natke, Donath, & Kalveram, 2003). Response magnitudes in a different speech condition ("*you know Nina?*", mean 31.5 ±18.7 cents), in turn, were larger compared to a single vowel condition (sustained /u/, mean 21.6 ±11.7 cents) (S. H. Chen et al., 2007). Response latencies were shorter in the speech condition (mean 122 ±63 ms) of S. H. Chen et al. (2007)) and the singing condition (142.1 ms) of Natke et al. (2003) compared to the vowel condition (mean 154 ±79 ms in S. H. Chen et al. (2007), and 154.6 ms in Natke et al. (2003)).

2.3.2 Pitch-shift Responses Are Optimal for Small Perturbations

To determine the most effective parameters for eliciting pitch-shift responses to pitch-shift stimuli, researchers have manipulated intensity, magnitude, latency, onset velocity, and duration of pitch-shift perturbations. It turns out that the intensity of the speakers' voice (65, 75, 85 dB SPL) and intensity of masking noise (none, 50, 60, 70 dB SPL) are not critical parameters affecting pitch shift responses (Burnett et al., 1998; Larson, 1998). Liu and Larson (2007) showed the percent compensation of the pitch-shift response magnitudes is greatest for a 10 cents stimulus magnitude (over 90% correction: pitch-shift response amplitude divided by pitch-shift stimulus magnitude) but compensation decreased as stimulus magnitudes increased (about 37% for 50 cents stimulus magnitudes; 20-40% for 100 cents in Hain et al. (2000)). This suggests that pitch-shift response is tuned to the correction of small pitch changes rather than large perturbations. One viable explanation for this difference is that smaller perturbations do not

appear to be deviant or sound like ‘alien voices’ but rather are correctible perturbations. Large perturbations that seem to be unlikely errors or ‘alien voices’ are less likely to be corrected.

In terms of stimulus latency (delayed by 0, 50, 100, 200, 300, 500 ms), there were no effects of pitch-shift stimulus latency on peak amplitude, latency, peak time, and onset of pitch-shift responses (Hain et al., 2001). In terms of stimulus onset velocity (10000, 1000, 500, 200, 100 cents/sec), pitch-shift response velocity increased with the increase of stimulus velocity (slope), whereas pitch-shift response latency, peak time, and amplitude showed an inverse relationship with stimulus velocity (Hain et al., 2001; Larson et al., 2000). The inverse relationship between response magnitude and stimulus velocity suggests that pitch-shift stimuli with lower velocity may be more effective in eliciting corrective responses than those with higher velocity.

2.3.3 Two Responses in the Pitch-shift Response

Increases in pitch-shift stimulus duration from 20 to 100 ms did not contribute to significant differences in pitch-shift responses (Burnett et al., 1998; Larson, 1998). However, as stimulus duration increased from 100 to 500 ms, voice F0 responses with longer duration and greater magnitude were elicited. However, these extended responses were composed of multiple response peaks, suggesting that there may be a secondary voice F0 response.

To examine the existence of multiple responses, Hain et al. (2000) and Larson (1998) asked the participants to either ignore the PSS, change voice F0 in the direction opposite to the PSS, or change voice F0 in the same direction as the PSS. The earliest vocal response (VR1) with a latency of 100-150 ms was highly resistive to voluntary modifications, whereas the later vocal response (VR2) at latencies of 250-600 ms could often be made in the instructed direction. This work suggested that VR1 is an early automatic response of the audio-vocal system that compensates for small perturbations in voice F0 feedback. VR2 seems to be a voluntary

mechanism that people use consciously to modulate their voice in speaking or singing.

2.3.4 Linguistic Specificity of Pitch-shift Responses: Experiments on Mandarin Language Production

Native Mandarin speakers show specific compensatory pitch-shift responses when producing Mandarin syllables. Their responses had shorter latencies than English speakers when the stimuli contain lexical tones: approximately 143 ms which is shorter than the 180 ms duration of a word (Jones & Munhall, 2000, 2002). In addition, Mandarin speakers produced larger amplitude in corrective responses for tones (3 bi-tonal patterns (T1-T1, T1-T2, T1-T4) of /ma ma/, mean 70±31 cents) than English speakers' pitch-shift response amplitude for the sustained vowel /a/ (mean about 30-40 cents and rarely exceeding 50 cents) (Burnett et al., 1998; Hain et al., 2000; Larson et al., 2001; Xu et al., 2004). However, the response magnitudes of Mandarin speakers were similar to non-Mandarin speakers in a singing condition (mean 66.1±30.4 cents) (Natke et al., 2003). This suggests that production of tones could require active monitoring of voice F0 similar to that of singing. Alternatively, linguistic experience may shape pitch-shift responses such that tones have more precise representations than productions without a specific tonal target. The conjecture that Mandarin speakers respond faster, possibly with larger amplitude, than English speakers suggests that native speakers of Mandarin specifically and differentially regulate their voice F0 by rapid adjustment of internalized pitch representations.

2.3.5 Long-term Adaptation of Pitch Responses

An independent but related paradigm has studied the effect on F0 when speaking under gradually altered pitch feedback. In these studies of sensorimotor adaptation there are typically three phases: 1) a baseline phase (10 utterances) where auditory feedback is normal; 2) a training

phase (120 utterances) where pitch is increased (Up condition) or decreased (Down condition) successively by 1 cent until the feedback received was one semitone (100 cents) above the subjects' true vocal pitch. Immediately following these 100 trials were 20 trials in which feedback was held at 100 cents above or below the subjects' original F0; 3) followed by a test phase (10 utterances) where participants receive normal auditory feedback. In Jones and Munhall (2002), ten female native speakers of Mandarin were asked to produce /ma/ with the Mandarin high level tone for 2 seconds. The results show that participants compensated for the pitch-shifted feedback. When feedback was suddenly returned to normal after the short-term exposure to the altered feedback, participants overcompensated and showed a negative aftereffect. These results paralleled those found for English speakers in Jones and Munhall (2000). Both Mandarin speakers and English speakers showed sensorimotor adaptation. When the subjects heard their feedback suddenly returned to normal (unmodified feedback in the test phase), the participants raised their pitch compared to the final 20 trials in the training phase of Up-shift condition, but lowered their pitch compared to the final 20 trials in the training phase of Down-shift condition.

Jones and Munahll (2005) examine the adaptation effect on another tone, i.e., rising tone (T2), in Mandarin. In their study, the pitch in the training phase was increased successively by 1 cent until the feedback received was one semitone (100 cents) above the subjects' true vocal pitch. For both tone 1 and tone 2, the participants' averaged voice frequencies were higher in the test phase than in the baseline. It suggests that the participants were adapted to the altered feedback. However, the adaptation effect persisted for tone 1 in the 20 utterances of the test phase, but decreased for tone 2 over the 20 utterances. The difference in decay rates of tone 1 and tone 2 suggests adaptation to gradually altered feedback is context dependent, which is important for Mandarin, as different tone contexts determine linguistic contrasts.

2.4 External Factors related to Pitch Processing

The pitch-shift paradigm reviewed above shows that speakers are able to rapidly adjust their speech when it does not fit the expected goal. The rapid adjustment provides a window to examine the expected speech motor command and begin to understand internalized pitch representations in the brain. It also provides tools to test whether internalized pitch representations can be reshaped and if adult second language learners can acquire new pitch representations.

Pitch-learning is expected to be affected by linguistic experience and extralinguistic experience such as musical training. This dissertation will examine the external factors by recruiting four different populations (naïve people, L2 learners, native speakers, and musicians). Another interesting external factor that will be considered is real-time visual feedback of pitch, which is suggested to be an advantageous tool for instructing singers (Howard et al., 2007; Thorpe, 2002; Wilson et al., 2008). In what follows, I will review the effects of language experience and musical experience on tone identification and discrimination, and the potential role of real-time visual feedback for motor control of pitch.

2.4.1 Language Experience Effect

The perturbation studies have shown that speakers' response magnitudes to perturbed speech may be dependent on speaker's language experience (Section 2.3.4). Native speakers of Mandarin have stronger compensation for tonal perturbations than native speakers of English did for vowel perturbations. Liu, Wang, et al. (2010) compared vocal responses to pitch perturbations between Cantonese speakers and Mandarin speakers and found that Cantonese speakers had smaller pitch-shift amplitude than Mandarin speakers for larger amplitude pitch-shift stimuli (±200 and ±500 cents). Similarly in Chen et al. (2012) which examined event-related potentials,

Cantonese speakers had larger P2 amplitudes to -200 cents and -500 cents stimuli than Mandarin speakers. Both studies suggest that the correction for pitch perturbation is subject to the specific tonal system of a language. The language-dependent modulation of vocal responses not only appears in F0 control but also in vowel formants (Mitsuya, Samson, Menard, & Munhall, 2013). French speakers and English speakers both showed compensatory production to formant shifts. However, French speakers responded to smaller pitch-shift stimuli (started compensating their F2 with a -162.11 Hz perturbation, compared to English speakers who started at a -260.59 Hz perturbation) and showed greater compensation than English speakers.

This language specificity also appears to be observed in studies of tone perception and the neural encoding of pitch. Some research shows that linguistic experience may facilitate tone perception (Y.-S. Lee, Vakoch, & Wurm, 1996; Wayland & Guion, 2004). For instance, native Mandarin speakers outperformed native English speakers in discriminating the Thai mid-tone and low-tone (Wayland & Guion, 2004). This study suggests that the perception ability of tone system may be transferred from one language to another language. However, other research argues that linguistic experience does not necessarily facilitate tone identification (Cooper & Wang, 2012; Francis, Ciocca, Ma, & Fenn, 2008). Francis *et al.* (2008) examined Cantonese tone identification by recruiting English and Mandarin listeners. The results show that English and Mandarin listeners did not differ significantly on the pretest performance of identifying individual Cantonese tones, but both groups had a significant improvement after training. However, the two groups differed in terms of the tones they found most difficult and easiest to identify, as they gave different feature weighting in the identification: English listeners tended to give more weight to Height than they did to Direction, while Mandarin listeners gave more weight to Direction than to Height. Cooper and Wang (2012) discovered the same pattern. No significant difference in accuracy on either the pre- or post-test was found between Thai and

English listeners identifying Cantonese tones. Thus, there is an evidence of native language influence (tone language or intonation language) from the perspective of feature weight, but whether the influence is facilitatory depends on the F0 patterns the listeners have been exposed to in their native language and the complexity of tone system the listeners would have to identify. It seems that the complexity of Cantonese tone system makes the identification difficult.

Even though behavioral research reviewed above does not consistently argue for the facilitation of linguistic experience on tone processing, brain research provides some evidence for the effect of language experience on tone perception. The mismatch negativity (MMN) evoked response to an oddball stimulus paradigm has been utilized to explore the cortical processing of linguistic pitch contours (Chandrasekaran et al., 2007b) and nonspeech stimuli (Chandrasekaran, Krishnan, & Gandour, 2007a, 2009b) for both Mandarin listeners and English listeners. Mandarin listeners showed larger MMN responses in the T1/T3 (standard/oddball) condition than the T2/T3 (standard/oddball) condition, whereas English listeners showed no significant difference between the two conditions (Chandrasekaran et al., 2007b). As in Francis *et al.* (2008), native speakers of Mandarin tended to focus on dynamic cues such as pitch direction and slope, while non-tonal English speakers attended to the cue of pitch height. It suggests that the early stage of pitch processing may be shaped by the relative saliency of acoustic dimensions encoded in a particular language. The effect of language experience on pitch processing also extends to certain nonspeech stimuli, as enhanced MMN responses were elicited for native speakers of Mandarin relative to English when the stimuli were nonspeech homologues of native pitch contours (T1/T2) (Chandrasekaran et al., 2007a, 2009b). A short period of tone training would be able to decrease the MMN amplitude differences between native speakers of English and native speakers of Mandarin (Kaan, Barkley, Bao, & Wayland, 2008). The English group became more similar to native Mandarin group in terms of the sensitivity to

F0 contours after repeated exposure.

Hemispheric lateralization studies of pitch processing also lend support to the effect of language experience. Non-lexical pitch perception is predominantly lateralized in the right hemisphere, for native speakers of Mandarin, while monitoring of lexical tones occurs predominantly in the left hemisphere (Van Lancker & Fromkin, 1973; Y. Wang et al., 2001). However, language training drives brain plasticity by recruiting additional cortical regions. Before training, right hemisphere activation was observed for non-tone speakers. After lexical tone training, language-related areas such as left superior temporal gyrus (Wernicke's area) were activated for non-tone speakers (Guenther et al., 1998; Yue Wang et al., 2003; Wong & Perrachione, 2007). This result suggests that the perceptual and neural systems involved in processing differences in pitch contours are still malleable, even in adulthood.

The experience-tuned pitch processing occurs at the brainstem level as well. The frequency following response (FFR) appears to originate in the auditory brainstem (inferior colliculus) and encodes the energy of the stimulus fundamental frequency. The FFR was recorded from the scalp as native speakers of Mandarin and native speakers of English listened to the four lexical tones in Mandarin (/yi1/ 'clothing', /yi2/ 'aunt', /yi3/ 'chair', /yi4/ 'easy') (Krishnan et al., 2005). The FFR pitch strength, as measured by average autocorrelation magnitude, was significantly greater for the Mandarin group than for the English group across all four Mandarin tones. The FFR pitch tracking accuracy, as measured by crosscorrelation between the F0 contours extracted from the original speech stimuli and those derived from the FFR waveforms, was more variable for the English group compared to the Mandarin group. This suggests that neural mechanisms are shaped by pitch contours that are specific to a language. Interestingly, the group difference disappeared when participants were asked to listen to the synthesized Mandarin monosyllable /i/ with linear rising and falling f0 ramps (Xu, Krishnan, & Gandour, 2006). It indicates that the

pitch encoding at the brainstem level is specific to pitch contours within the native listeners' experience.

2.4.2 Musical Experience Effect

In addition to language experience, musical experience is also relevant to tone processing, as musical note and lexical tone both have to do with learning pitch contours. Research shows that English-speaking musicians had better performance in Mandarin tone identification on intact, silent-center, and onset-only syllables than English-speaking nonmusicians (C.-Y. Lee & Hung, 2008). However, for the same task, Mandarin-speaking musicians did not outperform Mandarin-speaking nonmusicians (C.-Y. Lee & Lee, 2010). Musical experience facilitated tone word learning for listeners without a tone language background, but not for those whose native language is a tonal language (Cooper & Wang, 2012). It suggests that the combination of musical experience and linguistic experience may not be additive. Additionally, the insignificant correlation between absolute musical note identification task and Mandarin tone identification task reveals that distinct processing mechanisms may involve in linguistic and musical domains (C.-Y. Lee & Lee, 2010).

Musical training experience may provide an advantage in cortical (measured by MMN) or subcortical (measures by FFR) processing of linguistically relevant pitch contours. When participants listened to three randomly presented Mandarin resynthesized tones (/mi1/ 'to squint', /mi2/ 'bewilder' and /mi3/ 'rice'), musicians showed stronger FFR amplitudes than nonmusicians at the brainstem level (Wong & Perrachione, 2007). The musicians also did better at the tone identification and discrimination tasks. Their identification and discrimination scores were also highly correlated with FFR pitch tracking in tone 3. The findings suggest that musicians who have to be sensitive to note variations may have an enhanced ability to learn

lexical tones.

To compare the musicians' performances and native speakers' performances, native speakers of Mandarin were included in Chandrasekaran, Krishnan, and Gandour (2009a). Instead of using the synthesized tones, they employed three iterated rippled noise (IRN) stimuli to examine acoustic shifts in stimuli that fall between- or within-tonal categories. Two of the IRN stimuli represented curvilinear Mandarin tones (T1, T2). A third represented a linear rising ramp (T2L) that does not occur in Mandarin natural speech. The between-category contrast (T1/T2) and the within-category contrast (T2L/T2) gave rise to two passive oddball conditions. Native speakers of Mandarin showed larger MMN mean amplitude than either musicians or nonmusicians. Musicians in turn showed larger MMN mean amplitude than nonmusicians. The MMN mean amplitude of the between-category contrast (T1/T2) was significantly larger than the within-category contrast (T2L/T2). This suggests that experience-dependent plasticity of pitch processing is not domain-specific (language vs. music), but is sensitive to the long-term exposure to the context (native vs. nonnative, or professional vs. unprofessional). Thus, musical training experience could facilitate pitch processing not only in music but also in language.

The reverse transfer (from language to music) also takes place. Bidelman et al. (2011) asked participants listened to a lexical tone (Mandarin tone 2; T2) and a pitch interval (melodic major third; M3). Musicians and native speakers of Mandarin showed greater pitch strength in FFR irrespective of domain (language or music) than nonmusicians. However, in the comparison of musicians and native speakers of Mandarin, the pitch strength was greater for musicians across domains but only in a limited number of time frames. This confirms the previous research that the processing of pitch representation is experience-dependent. The brainstem processing of musicians and native speakers of Mandarin is tuned by long-term exposure to the pitch patterns inherent to a particular context. Their abilities in pitch decoding can be transferred across

domains. That suggests the pitch experience obtained in one domain may facilitate the learning capability in another domain.

2.4.3 Real-time Visual Feedback in Instruction

The last external factor of this dissertation is the potential role of real-time visual feedback for motor control of pitch. The effect of visual feedback has been explored in singing instruction (Howard et al., 2007; Thorpe, 2002; Wilson et al., 2008). In traditional instruction, oral feedback from the instructor is often delayed, as it is provided after the student's performance on singing. It depends on the student's ability of interpretation on what the teacher says and his memory of how the performance was produced (Welch, 1985). The only online feedback is auditory (and somatosensory) which is not shared by both instructor and learner. Wilson *et al*. (2008) assessed the effect of concurrent visual feedback on learner singers' (nonprofessional musicians) pitch-matching vocal abilities. The participants were divided into three groups, including two experimental groups and one control group. The experimental groups received different types of visual feedback on computer screen. In one group, the sung note was tracked by a pitch trace line; in another group, a sung note changed the color of piano key only when it was being sung.

The results of pre-test, training, and post-test show that the leaners who received visual feedback (irrespective of the mode), compared to the control group without any feedback, improved in their pitch accuracy at the post-test. It suggests that learners can have higher pitch accuracy with the real-time visual feedback than with traditional instruction only. Notice that there are many kinds of visual feedback that could be displayed to leaners: pitch trace, spectrogram, vocal tract area, or acoustic pressure waveform (Howard et al., 2007). One important factor which has been suggested is that the feedback information should be simple to interpret and relevant to the task (Thorpe, 2002). Visual feedback has not been explored in the

context of pitch-shift responses, but will be employed in this dissertation by providing visual feedback of fundamental frequency.

2.5 Summary

Tone learning is challenging for adult second language learners whose native languages are non-tonal. Even for young children whose native language is a tone language, it could take up to 10 years for their performance to reach an adult level. Most of the L2 tone learning research has focused on the ability to perceive tone categories and found that learners can make a considerable improvement after a period of training, among which an adaptive training program could facilitate the learning most. What is unclear is whether L2 learners can change their ways of controlling laryngeal motion and reshape the internalized pitch representations through exposure to tonal language training. The studies that explore the internal models by using pitch-shift paradigm have shown that people who speak tonal languages may have different pitch-shift responses from non-tone speakers, showing that their internalized representations for pitch are not the same. This dissertation will extend this research to investigate learners of Mandarin. Furthermore, there could be stimulus specificity of pitch-shift responses: different lexical tones may lead to different pitch-shift responses due to the inherent distinction in pitch height and movement direction of tones. In order to bridge the gap between perception (discrimination) and production (motor control) in second language tone learning research, non-/linguistic tone discrimination tasks and pitch-shift tasks with linguistic tone stimuli will be used to examine the correlation between the two. By making a connection between speech physiology and second language acquisition, the learners' capability of learning tonal languages will be better understood.

Linguistic background and musical background also affect pitch processing. Speaking a

tonal language may help one to learn another tonal language, though the extent of facilitative effect depends on the complexity of the native and target tone systems. Having musical training could aid in learning linguistic tones as well, which suggests that pitch experience in one (musical) domain may transfer to another (linguistic) domain. Providing real-time visual feedback of pitch trace has helped singer's performance but has not been tested for tone learning. I will test whether concurrent visual feedback can help speakers to control their voice pitch and to adjust their pitch-shift responses, and to investigate whether this feedback suppresses pitch-shift responses. This study is relevant to L2 tone learning, as the learners could potentially use real-time visual feedback to learn pitch contrasts and to correct tone errors. In order to explore the effects of linguistic and musical experience, I will recruit naïve people who were never exposed to tonal languages, adult L2 learners of Mandarin, musicians (trained vocalists), and native speakers of Mandarin.

CHAPTER 3
STUDY 1: TONE PERCEPTION AND SENSORIMOTOR RESPONSES TO SUSTAINED VOWELS

3.1 Introduction

Learning to speak a new language requires learning of a new motor skill. To produce new sounds and new sound combinations, a learner needs to blend the perception of new sensory information with new motor commands to achieve the desired result. Although important distinctions remain between skill learning and language learning, the common features of perceptual and motor learning may be mediated by common brain mechanisms. One common mechanism might be internal models that are thought to mediate sensory input and motor commands in order to achieve a certain movement goal (Guenther et al., 1998; Hickok et al., 2011; Jordan & Rumelhart, 1992; Kawato, 1999; Lalazar & Vaadia, 2008). In the complex case of vocalization for speech, the brain initiates a speech motor command that is accompanied by generation of a copy of the expected motor commands that is sent to sensory processing regions, i.e., efference copy. The brain actively compares the anticipated result of the efference copy with the desired result of the motor command. If there is a mismatch between the predicted result of the motor command and the desired outcome, a correction signal is generated to modify the motor commands.

Presumably, pitch control is very important for Mandarin tones. If there is a possibility that pitch could be disrupted easily (where mismatch happens), Mandarin tone is susceptible to considerable distortion through auditory feedback. This susceptibility was studied by Xu and colleagues who would like to know if Mandarin speakers are more or less susceptible (Liu, Xu, & Larson, 2009; Xu et al., 2004). However, they found inconsistent evidence. In one case, they found larger responses (Xu et al., 2004). In another case, they found smaller responses (Liu, Xu,

et al., 2009). Xu and colleagues never looked at L2 learners. We know that using pitch-shift paradigm to probe the stability of internal models is a general question in the field. Applying this to investigate the stability of internal models in L2 learning process is a new concept and a major contribution in this dissertation.

In Study 1, I will compare the perception and pitch-shift performances of three speaker groups: naïve speakers with no exposure to tonal languages, L2 learners of Mandarin, and native speakers of Mandarin. Three tasks are administered to all the three groups. First, in the pitch-shift task, we used pitch-shift responses to the simple vowel /a/ to examine tonal language experience effect on audio-vocal interactions in a broader spectrum of language experience (from naïve to expert). If Mandarin speakers show differential responses, it provides evidence that linguistic experience shapes an internal model of tone production. Correspondingly, if L2 learners show pitch-shift responses that are more similar to the native Mandarin speakers than naïve speakers, it would suggest that even some tonal-language experience can alter sensorimotor relationships in adult L2 learners. In the longer term, examining production by using pitch perturbations may enable prediction of individual aptitude for Mandarin tone learning. Mandarin speakers may have developed internal models of tone production that are controlled on a syllabic basis. If their system is perturbed, Mandarin speakers may respond with a different F0 contour in the temporal domain reflecting the influence of their internal tone models whereas non-tone speakers may respond on a different scale or direction.

It is equally important to study how potential reshaping of vocal control is related to the perceptual learning that is also presumably taking place. Thus, a musical/nonlinguistic tone discrimination test and a Mandarin tone discrimination test were included in addition to the pitch-shift task. Looking at tone perception in addition to audio-vocal responses will contribute to a more complete picture of how internal models for tone proficiency are organized. It is

expected that native speakers of Mandarin and L2 learners may have better tone discrimination ability than naïve speakers, and that their perception ability should be correlated with their voice F0 control ability.

In summary, I am asking the following questions: i) whether Mandarin tone discrimination ability is related to tonal language exposure by comparing Mandarin speakers, L2 learners, and naïve speakers without tonal language experience; ii) whether the ability to discriminate Mandarin tones is related to the ability to discriminate nonlinguistic tones in the three groups; iii) whether changes in sensorimotor control due to tonal language experience are evidenced by differences in pitch-shift responses between the three groups. In this study, I introduce a methodology for comparing changes in F0 contour shape (Kosling, Kunter, Baayen, & Plag, 2013) following perturbation, along with the typical approach that compares peak amplitude and peak latency. I hypothesize that patterns of tone discrimination and sensorimotor interactions in L2 learners who may have started to reshape their internal models will resemble native speakers of Mandarin, which, in turn, will contrast with the patterns of naïve speakers. I also hypothesize that perception and production measurements are required in order to differentiate the scope of language learning influences.

3.2 Experiments in Study 1

The three tasks conducted in Study 1 are presented below. The pitch-shift task investigates the language experience effect on sensorimotor control over voice F0. The nonlinguistic tone discrimination task examines the participants' ability to discriminate two tones with different fundamental frequency. The Mandarin tone discrimination task explores the participants' ability to differentiate lexical tone differences. The same participants participated in all the three tasks.

3.2.1 Participants

Twenty-nine participants participated in the experiment including the following groups: Ten naïve speakers (10 female) who were never exposed to tonal languages and whose native language is English, 10 adult L2 learners of Mandarin whose native languages is English (5 male; 3 first-year students, 5 second-year students, 2 third-year students), and 9 native speakers of Mandarin (4 male). All were students at University of Illinois at Urbana-Champaign (age: 20-30 years old). Based on the post-hoc interviews, the average period of musical training (instruments) before college was 6.5 ± 1.3 years for naïve speakers (from the report of 8 participants), 6.3 ± 1.4 years for L2 learners (from the report of 9 participants), and 4.6 ± 1.2 years for Mandarin speakers (from the report of 8 participants). None of them reported a continuation of musical training after college. All participants passed a hearing screening at 20 dB HL bilaterally at 125, 250, 500, 750, 1000, 2000, 3000, and 4000 Hz. None of the participants reported a history of neurological or communication disorders. They were paid $10 for their participation. All methods reported herein were approved by the Institutional Review Board at the University of Illinois and all participants provided written informed consent.

3.2.2 Pitch-shift Task

3.2.2.1 Materials

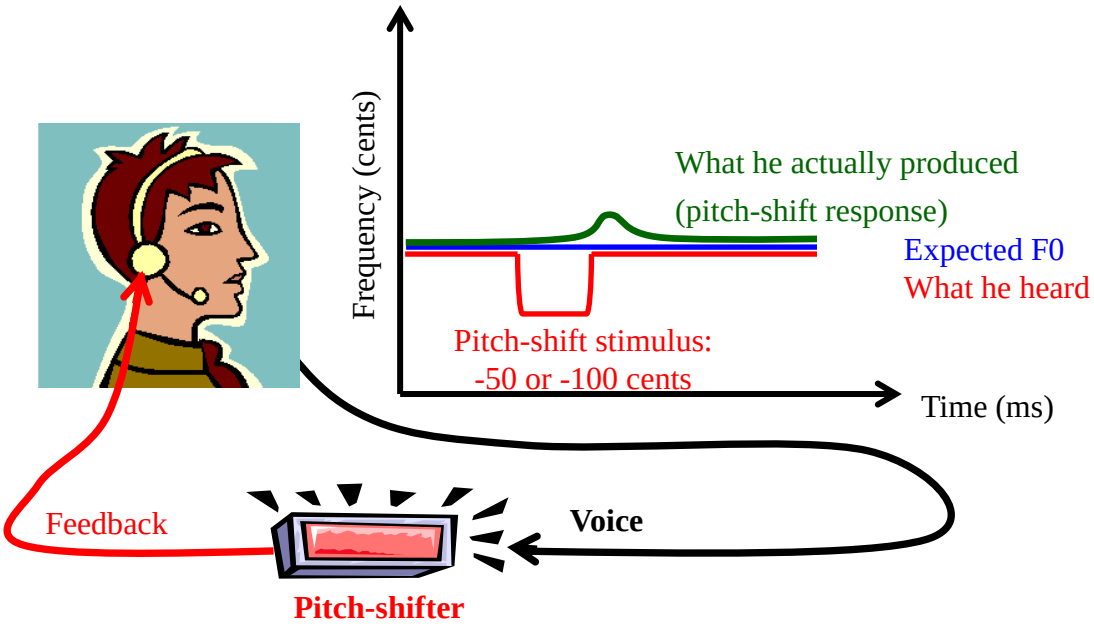
On each trial, participants first heard a male /a/ vowel recording which lasted for 3 s. Then they repeated the /a/ vocalization for 5 s after the vowel stimulus ended. They were instructed to hold vocal pitch and volume as steady as possible and to ignore any changes in what they heard. At a randomized interval (200, 300, and 400 ms) after vocalization onset, the pitch of the feedback signal was altered for 200 ms by either a ±50 cent shift (blocks 1 and 2) or a ±100 cent shift (blocks 3 and 4) with shift direction (upward or downward) randomized. Splitting the trials

into blocks 1 and 2 or into blocks 3 and 4 was done for participants to take a break. The inter-stimulus interval was fixed for all blocks (1000 ms). Thirty-two vocalizations were recorded per block with two pitch-shift perturbations per trial for a total of 256 perturbations with the number of upward and downward shifts approximately equal.

3.2.2.2 Procedures

The pitch-shift paradigm is illustrated in Figure 3.1. The participant is instructed to produce /a/ with a steady vocal pitch and volume. What he is supposed to hear is his steady pitch (the blue line). However, we manipulate the sound by decreasing or increasing (decreasing in this figure) the pitch by 50 or 100 cents, and feed that shifted sound back to his ears (the red line). Typically, people would produce a compensatory pitch-shift response, which is in an opposition direction to the pitch-shift stimulus (the green line).

Figure 3.1 Illustration of the pitch-shift paradigm.



During the entire experiment, participants were seated in a sound booth and wore Sennheiser HD-280 Pro headphones and a headworn Shure microphone placed 2 cm away from the corner of the mouth. They were trained to produce the vowel /a/ at approximately 70 dB sound pressure level (SPL) while self-monitoring their vocal volume on a Dorrough Loudness Monitor (model 40-A) placed in front of them. The voice signal from the microphone was amplified with a YAMAHA mixer (MG102c) and sent to an Eventide Ultra-Harmonizer (model H7600) that generated pitch shifts. The participant's own voice signal was amplified with a Samson S-phone Headphone Amplifier and played back to him/her at approximately 80 dB SPL to reduce the possible influence of bone conduction. MIDI software (MAX/MSP v.5 by Cycling 74, Walnut, CA) connected to a MOTU (model UltraLite-mk3 Hybrid) controlled the timing, duration and magnitude of the pitch shifts via the Eventide ultra-harmonizer. The participant's voice, altered feedback signal, and pitch-shift events (transistor transistor logic (TTL) pulses) were digitized at 5 kHz per channel with WINDAQ/Pro software (DATAQ Instruments, Inc., Akron, OH).

3.2.3 Nonlinguistic Tone Discrimination Task

3.2.3.1 Materials

The nonlinguistic tone discrimination task examines the participants' ability to differentiate the pairs of tones that differ in fundamental frequency. To estimate nonlinguistic tone discrimination, the Adaptive Pitch Test (Mandell, 2009) on the tonometric website was presented via a laptop.

3.2.3.2 Procedures

The volume was adjusted to each subject's most comfortable level. Participants listened to a

series of two short tones and were asked whether the second tone was lower or higher in pitch than the first tone. After clicking on the start button, the program starts by presenting a tone pair with 96 Hz pitch difference. Participants could replay the sound as many times as needed. Once the subject made their choice by clicking the “higher” or “lower” button, the next pair of tones was then immediately played. This adaptive software either reduced the pitch difference for correct responses or increased the difference for incorrect responses. After determining the smallest tone difference in Hertz that the subject could consistently discriminate, the difference in Hertz was presented on the webpage.

Based on results from 11,000 persons, less than 0.75 Hz indicates exceptional discrimination, less than 1.5 Hz indicates very good discrimination, less than 6 Hz indicates normal discrimination, less than 12 Hz indicates low-normal discrimination, and greater than 16 Hz indicates a possible pitch perception deficit. Participants repeated this task twice (i.e., 2 attempts) to obtain a reliable estimate of pitch discrimination. Each attempt took about 3 minutes to complete.

3.2.4 Mandarin Tone Discrimination Tasks (MD1 & MD2)

3.2.4.1 Materials

In the Mandarin tone discrimination task, participants listened to pairs of Mandarin words with identical segments but possibly differed in tone (tone 1: high level tone, tone 2: rising tone, tone 3: falling-rising tone, and tone 4: falling tone). They had to judge whether the tones are the same or different.

Given that there are four lexical tones in Mandarin, there were 16 combinations in the pairs of Mandarin words (Table 3.1). In order to make the “same” and “difference” trials to have equal probability of occurrence, word pairs with the same tone were weighted 3 more times than word

pairs with different tones.

Table 3.1. Sixteen combinations of the tone pairs in Mandarin

	Word 1	**Word 2**
Same tone	Tone 1	Tone 1
(4 possible	Tone 2	Tone 2
combinations)	Tone 3	Tone 3
	Tone 4	Tone 4
Different tones	Tone 1	Tone 2
(12 possible	Tone 1	Tone 3
combinations)	Tone 1	Tone 4
	Tone 2	Tone 1
	Tone 2	Tone 3
	Tone 2	Tone 4
	Tone 3	Tone 1
	Tone 3	Tone 2
	Tone 3	Tone 4
	Tone 4	Tone 1
	Tone 4	Tone 2
	Tone 4	Tone 3

The task was subdivided into 2 sections based on the voices. In the first section (MD1 for short), 56 trials ((4 pairs of "same tone" * 4 repetitions + 12 pairs of "different tones" * 1 repetition) * 2 words) were used and the two Mandarin words in each trial were spoken by the same female. The words in the first section consisted of /waŋ/ and /tɕy/. In the second section (MD2 for short), 84 trials ((4 pairs of "same tone" * 4 repetitions + 12 pairs of "different tones" * 1 repetition) * 3 words) were used and the two Mandarin words in each trial were spoken by two different females. The words in the second section consisted of syllables /ma/, /ʂu/ and /toŋ/. The second section was expected to be more difficult for listeners to discriminate than the first section because in the second section they had to ignore individual differences in pitch and

retrieve abstract representations for tone. Using multiple words in either MD1 or MD2 was to increase the total number of trials for the 16 combinations of word pairs. Using different Mandarin words in the two sections was also intended to avoid any learning effects.

3.2.4.2 Procedures

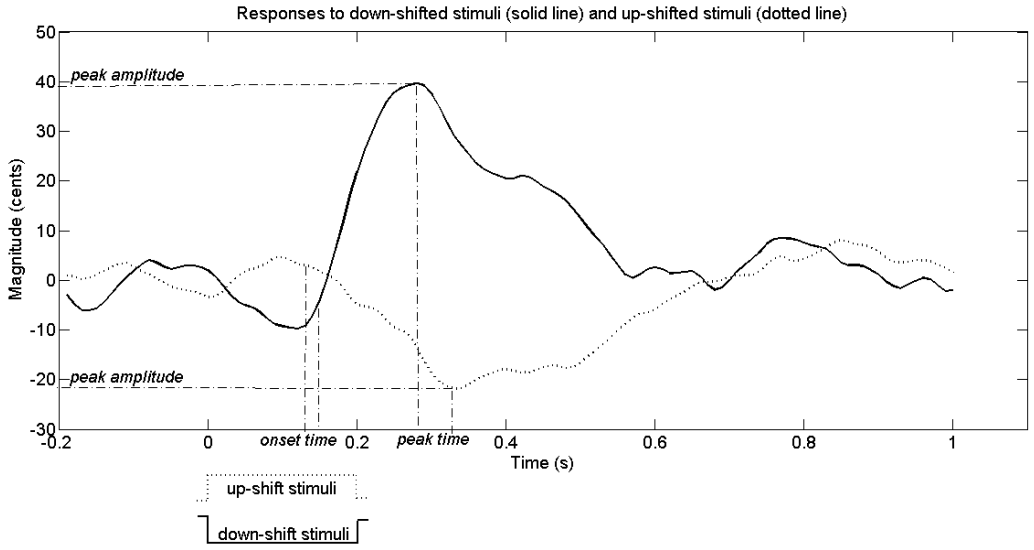
E-Prime software (v.1, Psychology Software Tools, Inc., Sharpsburg, PA) was utilized to present pairs of Mandarin tones that participants were required to discriminate. Participants were asked to judge whether the two words had the same or different tones by pressing a "same" or "different" button. As per the nonlinguistic tone discrimination task, participants could replay the sound as many times as they need. Before MD1 and MD2 were presented, participants were given a short practice session (3 trials). The overall accuracy of each section was presented on the screen at the end. The order of the tasks was counterbalanced across participants.

3.2.5 Data Analysis

In the pitch-shift task, the raw signals in WINDAQ were imported into MATLAB (R2009b, The Mathworks Inc., Natick, MA) and then sorted into files based on the direction of the pitch shift stimuli (up-shift vs. down-shift) and shift magnitude. For each vocalization, a 1.2 second window including a 200 ms pre-pulse baseline, the 200 ms pulse, and an 800 ms post-pulse period was selected for acoustic analysis. The *pitch values* were tracked using an autocorrelation algorithm in Praat (10 ms interval) and then converted into cents (cents = $1200*\log_2(F0/\text{baseline F0})$), where the baseline F0 is the mean F0 between 0-200 ms. Pitch-shift studies have shown that speakers may adjust their pitch in an opposite direction to pitch-shift stimuli (compensatory responses) but a certain number of responses may follow the direction of pitch-shift stimuli (following responses). Following responses and responses without steady baseline (the phase

before pitch-shift occurs) were excluded in this study so that only compensatory responses are considered. Then the onset time, peak time, and peak amplitude of the average responses were obtained in the following way (Figure 3.2): *onset time* was determined by a 10% increment from the baseline of averaged response; *peak time* and *peak amplitude*, which corresponded to the first peak of averaged response, usually occurred between 200-600 ms after the onset of the pitch-shift stimulus and were selected with a peak-picking algorithm in MATLAB.

Figure 3.2 Illustration of the onset and the peak of pitch-shift responses.



Note. The x-axis represents the time in second. The onset of pitch-shift stimuli is at time 0 second and the offset is at time 0.2 second. The y-axis represents the magnitude of pitch-shift responses measured in cents. The solid contours are compensatory responses to downward pitch perturbation. The dotted contours are compensatory responses to upward pitch perturbation.

To evaluate pitch-shift responses, a repeated measures ANOVA comparing the factors of GROUP (naïve speakers, L2 learners, and Mandarin speakers), DIRECTION (up-shift and down-shift), and MAGNITUDE (50 cents and 100 cents) was conducted separately for the onset time, peak time, and peak amplitude variables. Second, to evaluate nonlinguistic tone discrimination, linear models of GROUP x ATTEMPT were fitted with the perception data. Third,

for Mandarin tone discrimination, the data were binary-coded (either correct or wrong for each trial), so a logistic regression of accuracy was conducted as an appropriate approach for examining the main effects of GROUP and SECTION along with their interaction. Fourth, correlations were conducted to examine whether nonlinguistic tone discrimination performances were related to Mandarin tone discrimination performances.

Discriminant analyses were performed to examine how accurately the non-/linguistic tone discrimination scores and the pitch-shift responses could classify individuals into three language groups (naïve speakers, L2 learners, and Mandarin speakers). For the pitch-shift responses, the measurements involved onset time, peak time, and peak amplitude of averaged F0 contours per subject per condition. The goal of the discriminant analysis is to find a linear combination of the measures that best classifies the three language groups. Backward automatic stepping was used so that the highly correlated (i.e., redundant) variables were removed from the model.

In addition to the traditional way of measuring pitch-shift responses (onset time, peak time, and peak amplitude), I modeled the response of the whole F0 contours to the pitch-shift perturbation by using Generalized additive models GAMs (Wood, 2006). This longer term view of the F0 contour captures the distribution of entire F0 response that a single measure of amplitude might not include. GAMs have been used to model nonlinearity such as acoustic pitch measurements (Kosling et al., 2013), where they investigated the F0 contours of right-/left-branching compound words. Smooth functions in GAMs can model time-varying contours in the plane by predictor and response variable. My interest is how F0 responses change over time for the three GROUPs (naïve speakers, L2 learners, and Mandarin speakers), the two pitch-shift DIRECTIONs (up-shift and down-shift), and the shift MAGNITUDEs (50 cents and 100 cents). I used the mgcv package (Wood, 2011) for R (R core team 2012). A series of model comparisons was conducted to investigate whether inclusion of a predictor (independent variable)

leads to a significantly better fit of the model to the absolute F0 records (dependent variable). The F0 records fitted into the GAM were the *pitch values* obtained from the pitch tracking algorithm in Praat at 10 ms interval which were then converted into cents. First, SUBJECT and TIME treated as random effects, `s(SUBJ,bs="re")` and `s(TIME,bs="re")`, were included in the model, which serves as a baseline. Second, the GROUP of the speaker was included as a predictor. As the group differences are expected primarily over time, I included separate temporal smooths with restricted cubic splines `s()` for GROUP. Third, DIRECTION was included as a predictor, in order to see if speakers responded differently for upward and downward shifts. Temporal smoothing with restricted cubic splines for DIRECTION was included. Fourth, MAGNITUDE was included as a predictor, as pitch-shift responses have been found to be optimal for smaller perturbations (Liu & Larson, 2007). Temporally smoothed contours with restricted cubic splines for MAGNITUDE were included to explore whether the effect of MAGNITUDE may vary over time. Significance of parametric terms (GROUP, DIRECTION, and MAGNITUDE) is evaluated by means of the t-tests, while significance of nonparametric terms (smooth terms) is evaluated by means of Bayesian *p*-values.

3.3 Results

Since the perception data and production data violated the assumptions of normal distribution and homogeneity of variance, permutation tests were used to examine statistical significance. Permutation tests are a type of widely-used non-parametric tests. They use random shuffles of the observations to get the correct distribution of a test statistic under a null hypothesis. An *F* value is then calculated as done in ANOVA. The permutation process repeats many times (say 5000 times). The *p*-value under the null hypothesis is calculated by the percentage of repetitions in which the resampled *F* exceeded the *F* obtained from the original data.

3.3.1 Tone Discrimination

The means and standard deviations of the nonlinguistic tone discrimination task (TD) and the Mandarin tone discrimination task (MD) are depicted in Figure 3.3 and Figure 3.4, respectively. In the nonlinguistic tone discrimination task, a lower score in Hz corresponds to better discrimination ability. Descriptively, the Mandarin group had better nonlinguistic tone discrimination scores than the L2 learner group for both attempts, and the L2 learner group had better nonlinguistic tone discrimination scores than the naïve group for both attempts. There was an improvement from TD1 to TD2 for the naïve group and L2 learner group on the second attempt (TD2). The naïve group improved to the level of learners on their first attempt (TD1) while the L2 learner group improved to the TD1 level of Mandarin speakers.

Figure 3.3 Discrimination performance in the nonlinguistic tone discrimination task (TD) by group and attempt.

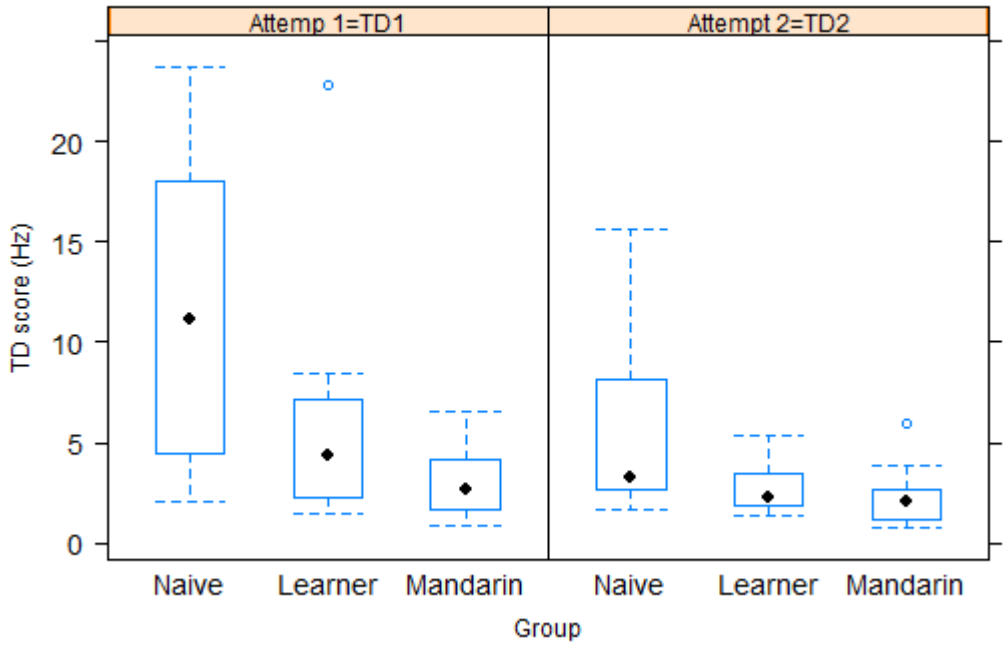
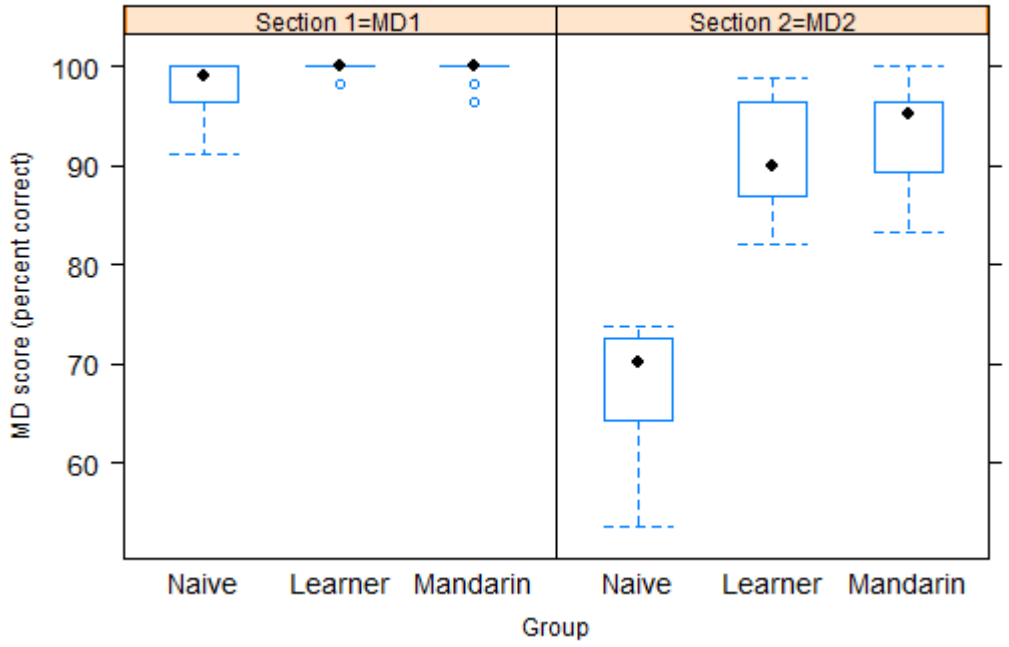


Figure 3.4 Discrimination performance in the Mandarin tone discrimination tasks by group and section (Section 1 (MD1) with the same speaker voice & Section 2 (MD2) with two speaker voices).



The permutation tests with two-way ANOVA show a significant main effect of GROUP ($F(2,52) = 165.096$, $p < .001$), and a significant main effect of ATTEMPT ($F(1,52) = 164.020$, $p < .05$) on nonlinguistic tone discrimination, but the interaction was not significant ($F(2,52) = 27.462$, $p = .306$). Post-hoc pairwise comparisons using a Wilcoxon rank sum test shows that the Mandarin group had significantly lower frequency discrimination intervals on both attempts compared to the naïve group ($p < .01$) (**MANDARIN>NAÏVE**), while no differences were found between the Mandarin group and the L2 learner group (**MANDARIN=L2 LEARNER**). The L2 learner group did not differ statistically from the naïve group (**L2 LEARNER=NAÏVE**). All listeners showed significantly better performances on the second attempt than the first attempt (**TD2>TD1**).

For the Mandarin tone discrimination task, a higher score (percent correct) corresponds to

better discrimination ability. The Mandarin discrimination task with two female voices (MD2) was more challenging than the task with a single female voice (MD1). Not surprisingly, the Mandarin group displayed better Mandarin tone discrimination scores than the L2 learner group, irrespective of the section (MD1 or MD2). Also as might be expected, the L2 learner group had better Mandarin tone discrimination scores than the naïve group, irrespective of the section. There was a marked decrease in percent accuracy from MD1 to MD2 for the naïve and L2 learner groups.

For Mandarin tone discrimination, a binary logistic regression was conducted to regress the accuracy of each trial (1 for correct and 0 for incorrect) on GROUP and SECTION (MD1 or MD2) along with interactions. The full model shows that there was no significant interaction between GROUP and SECTION (*Wald* $\chi^2(2) = 0.91$, $p = 0.636$). The reduced/simplified model generated by removing the interaction term shows that there was a significant main effect of GROUP (*Wald* $\chi^2(2) = 218.43$, $p < .001$) and SECTION (*Wald* $\chi^2(1) = 153.43$, $p < .001$). Pairwise comparisons show that the L2 learner group did not perform statistically differently from the Mandarin group ($p = .094$), while both the L2 learner group and the Mandarin group had better discrimination than the naïve group ($p < .001$ for both) (**MANDARIN=L2 LEARNER>NAÏVE** for the MD performances). The significant result for SECTION indicates the listeners had better performances on MD1 than MD2 (**MD1>MD2**).

To further investigate whether performances in the nonlinguistic tone discrimination task were correlated with the Mandarin tone discrimination task, distribution-free spearman correlation analyses between the TD scores and the MD overall accuracy were conducted. TD1 was positively and significantly correlated with TD2 ($r = .575$, $p < .01$) and TD2 was negatively correlated with MD1 ($r = -.567$, $p < .01$) but no other correlations were significant. The average of TD1 and TD2 scores was not significantly correlated with MD1($r = -0.121$, $p = .532$) or MD2

(r = -0.345, p = .067).

3.3.2 Pitch-shift Task

Only compensatory responses (opposite direction to the stimulus) with steady baselines were included in the data analysis, excluding 33% of the trials overall which were composed of either following responses (same direction to the stimulus) or non-responses (with no steady baseline). In any case, following responses constituted a minority of responses in each participant.

To investigate whether pitch-shift responses are affected by language experience, I conducted permutation tests with repeated measures ANOVAs comparing the factors of GROUP, DIRECTION, and MAGNITUDE for each dependent variable (response *onset time* (ms), *peak time* (ms), and *peak amplitude* (cents)). For response onset time, there were no main effects of GROUP (F(2,111) = .696, p = .501), DIRECTION (F(1,111) = 3.367, p = .069) or MAGNITUDE (F(1,111) = .099, p = .75). For response peak time, there were no main effects of GROUP (F(2,111) = 1.623, p = .202), DIRECTION (F(1,111) = .101, p = .752) or MAGNITUDE (F(1,111) = .364, p = .547). For response peak amplitude, there were no main effects of GROUP (F(2,111) = 0.701, p = .498), DIRECTION (F(1,111) = .182, p = .670) or MAGNITUDE (F(1,111) = 3.381, p = .069).

3.3.3 Discriminant Analyses: Classifying Speakers by the Use of Tone Perception Performances and Pitch-shift Responses

As a broader analysis of how the groups may differ when both perception and sensorimotor variables are considered, I conducted discriminant analyses. A backward automatic stepwise discriminant analysis was performed to find a linear combination of the measures (TD1, TD2,

MD1, MD2, onset time, peak time, and peak amplitude) that best classifies the three language groups (naïve speakers, L2 learners, and Mandarin speakers). Backward automatic stepping was used because there might be correlations between the variables (TDs, MDs, and audio-vocal responses). Variables with a probability of *F*-to-enter statistics smaller than the enter probability 0.1 were entered into the model if tolerance (the correlation of a candidate variable with the variables included in the model) permits. Variables with probability of *F*-to-remove statistics larger than the remove probability of 0.15 were removed from the model. *F*-to-remove statistics corresponds to the relative importance of variables with low tolerance meaning the variable may be redundant or highly correlated with another variable.

The stepping summary is displayed in Table 3.2. The first variable that was removed from the model was the onset time of down-shift 50 cents condition, since it had the lowest *F*-to-remove value that passes the remove limit of probability. The final model shown in Table 3.3 lists Wilk's Lambda which tests dispersion among all the groups on all the variables. The difference between each pair of groups on the six variables was significant (*Lambda* = 0.05; Approximate *F*-ratio = 12.131; *df* = 12,42; $p < .001$). The *F*-to-remove statistics and tolerance in Table 3.3 indicates that MD2 (large *F*-to-remove statistics and tolerance) contributed most to the model in predicting the group of an individual.

Table 3.2. Stepping summary of the Canonical Discriminant Analysis

	F-to-remove	Wilks's Lambda	Approx. F-Ratio	p-Value
DN50ONSET	-0.003	0.028	3.994	0.000
DN100PKTM	-0.154	0.029	4.566	0.000
UP100ONSET	-0.137	0.029	5.229	0.000
TD2	-0.316	0.030	5.910	0.000
MD1	-0.334	0.032	6.697	0.000
UP50PKAM	-0.747	0.035	7.412	0.000
UP50ONSET	-0.681	0.038	8.314	0.000
TD1	-0.978	0.042	9.258	0.000
UP100PKTM	-1.155	0.047	10.357	0.000
DN100ONSET	-0.724	0.050	12.131	0.000

Table 3.3. Variables included (left) and excluded (right) in the final classification model

Variable	F-to-remove	Tolerance	Variable	F-to-Enter	Tolerance
MD2	90.890	0.443	**TD1**	0.375	0.653
UP50PKTM	6.729	0.537	**TD2**	0.511	0.679
DN50PKTM	2.510	0.864	**MD1**	0.121	0.690
DN50PKAM	4.206	0.557	**UP50ONSET**	0.625	0.689
UP100PKAM	12.860	0.263	**UP50PKAM**	0.459	0.697
DN100PKAM	11.204	0.287	**DN50ONSET**	0.032	0.800
			UP100ONSET	0.617	0.976
			UP100PKAM	0.470	0.902
			DN100ONSET	0.724	0.712
			DN100PKTM	0.012	0.816

The classification matrix based on the classification functions in the final model is presented in Table 3.4a. Values in the diagonal of the classification table reflect the correct classification of individuals into groups. Figure 3.5a is a scatter plot of the individuals' canonical variable scores on two discriminant dimensions, indicating predicted group classification. The final model with six variables (including MD2 and other pitch-shift responses) had success in separating the three groups as shown by an overall classification accuracy of 97 % (Table 3.4a).

Table 3.4. Canonical Discriminant Analysis

a. Classification matrix and Jackknifed classification matrix based on the final model (cases in row categories classified into columns)

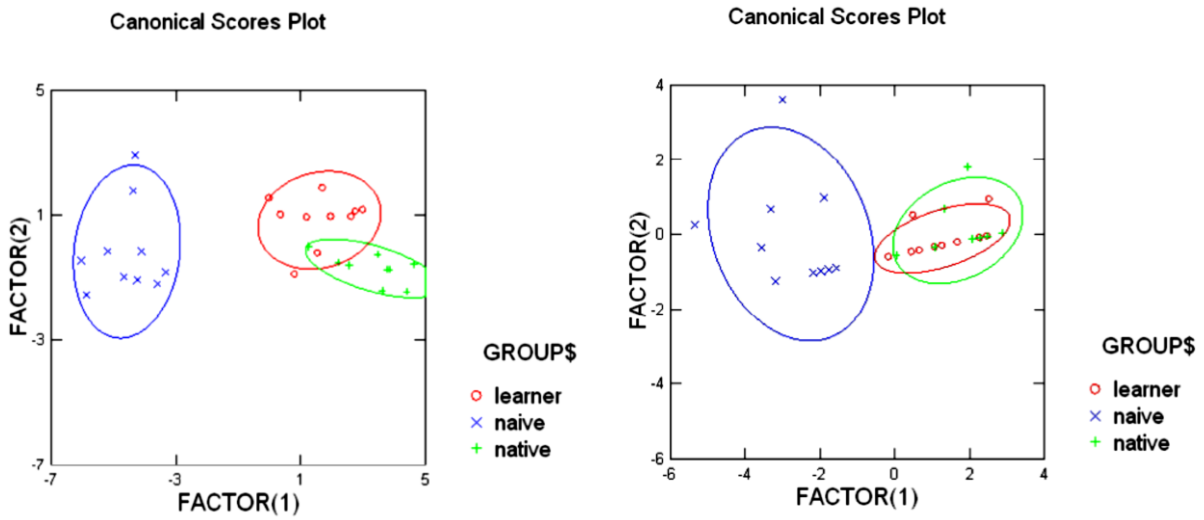
	Naïve	L2 Learner	Native	% Correct
Naïve	10	0	0	100
L2 Learner	0	10	0	100
Native	0	1	8	89
Total	10	11	8	97

b. Classification matrix by using MD1 and MD2 (cases in row categories classified into columns)

	Naïve	L2 Learner	Native	% Correct
Naïve	10	0	0	100
L2 Learner	0	6	4	60
Native	0	3	6	67
Total	10	9	10	76

Figure 3.5 ***a.*** Canonical scores plot by using the six predicting variables in the final model. ***b.*** Canonical scores plot by using MD1 (with the same speaker voice) and MD2 (with two different speaker voices).

a. b.



Note. Factor 1 is the first canonical discriminant function and Factor 2 is the second canonical discriminant function, which are linear combinations of the six included variables weighted by their discriminant function coefficients.

If only linguistic discrimination scores (MD1 and MD2) are included in the discriminant

analysis, then classification accuracy drops to 76% (Table 3.4b). Although individuals with tonal language experience (Mandarin and L2 learner) are clearly separated from those without (naïve), Mandarin speakers cannot be separated from L2 learners (Figure 3.5b). The comparison of Figure 3.5a and Figure 3.5b shows that the inclusion of pitch-shift responses among the predictor variables is needed to discriminate L2 learners from Mandarin speakers.

3.3.4 Generalized Additive Models: Modeling the Pitch Values in the Pitch-shift Task

In typical analyses of pitch-shift responses, two points on the F0 contours are selected, that include the onset and peak. However, there are many points on the F0 contours that could show meaningful response variation. A method is needed to examine if the temporal profile of entire F0 contours diverge according to group (Mandarin, L2, or naïve speakers), which is possible with Generalized Additive Models (GAM). This novel GAM analysis, adapted from methods reported by Kosling *et al*. (2013 and Sakai (2014), herein suggests that pitch perturbations lead to group differences in the temporal profile of F0 contours.

Generalized additive models (GAMs) were fitted with the absolute F0 values as the dependent variable. The restricted cubic splines smoother produces a smoothed generalization of the relationship between time and F0 values in the scatterplot .GAM analyses show that there were no significant differences in the F0 contours before perturbation onset (200 ms long). To examine how speakers respond to pitch-shift stimuli, I only considered the F0 contours after the onset of perturbation (1 second long for each curve) in GAMs. The sequence of model comparisons is summarized in Table 3.5. Each row compares two models, where the second model has one more predictor or interaction term than the first model. The evaluation is based on whether there is a reduction in deviation (Akaike Information Criterion, AIC for short), and whether this reduction is significant given the effective degrees of freedom (*edf*). Significance

was evaluated with an F test. The baseline model, which is not shown in the table, includes SUBJECT and TIME as random effects. The first row indicates that including the predictor GROUP reduced the AIC by -0.228. The *F*-test shows that inclusion of GROUP as a predictor did not lead to a significantly better fit of the model to the data (p=0.186). However, significant differences between the groups developed as time progressed (reduction in AIC 23.487, p<.0001). The model was improved when the predictors DIRECTION and MAGNITUDE, and their interactions with TIME were included, which significantly reduced the AIC.

Table 3.5. Generalized Additive Model: Sequential model comparison in Study 1

Predictor	*edf*	Reduction AIC	*F*	*p*
GROUP	2	-0.228	1.205	0.186
s(TIME, GROUP)	7.756	23.487	5.268	<.0001
DIRECTION	1	7304.016	2459.100	<.0001
s(TIME, DIRECTION)	16.872	2983	188.620	<.0001
MAGNITUDE	1	40.551	41.879	<.0001
s(TIME, MAGNITUDE)	7.597	35.350	6.997	<.0001

The final model presented in Table 3.6 indicates the GROUP, DIRECTION, and MAGNITUDE had significant influences on F0 changes over time. Since the interactions between TIME and GROUP, between TIME and DIRECTION, and between TIME and MAGNITUDE were all significant, stratified post-hoc comparisons were performed on the coefficients of 3 GROUPs, 2 DIRECTIONs and 2 MAGNITUDEs. Significance was evaluated with a Wald Chi-squared test for model coefficients. For GROUP, there were no significant differences in F0 contours between the Mandarin and the L2 learner groups ($\chi^2(1)=0.047$, p=0.83). However, the differences between the naïve and Mandarin groups ($\chi^2(1)=10.8$, p<.01) and between the naïve and the L2 learner groups ($\chi^2(1)=9.8$, p<.01) were significant. As for DIRECTION, the responses to upward shifts were significantly from the responses to downward shifts ($\chi^2(1)=102.1$, p<0.001). As for MAGNITUDE, the responses to pitch-shifted stimuli of 50 cents were significantly different from

the responses to pitch-shifted stimuli of 100 cents ($\chi^2(1)=7.5$, $p<0.01$).

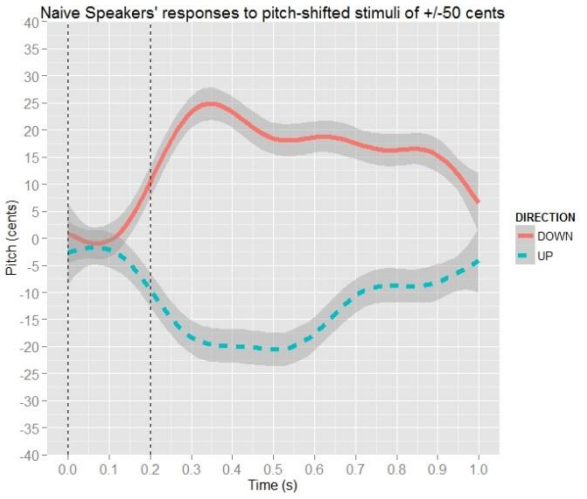
Table 3.6. Summary of final Generalized Additive Model in Study 1

Parametric coefficients				
	Estimate	Std. Error	*t*	*p*
intercept	12.405	1.276	9.723	<.0001
GROUPMandarin	1.076	1.798	0.599	0.549
GROUPNaive	0.855	1.750	0.488	0.625
DIRECTIONUp	-25.992	0.227	-114.700	<.0001
MAGNITUDE100	-1.479	0.227	-6.525	<.0001
Approximate significance of smooth terms				
	edf	Ref.*df*	*F*	*p*
s(SUBJ)	25.370	26.000	37.604	<.0001
s(TIME)	1.169e-15	1.000	0.000	0.953
s(TIME):GROUPL2	6.310	7.404	5.016	<.0001
s(TIME):GROUPMandarin	0.750	0.750	0.778	0.445
s(TIME):GROUPNaive	4.066	5.039	5.294	<.0001
s(TIME):DIRECTIONDown	8.503	8.648	43.644	<.0001
s(TIME):DIRECTIONUp	7.966	8.454	31.139	<.0001
s(TIME):MAGNITUD50	6.832	7.810	3.158	<.01
s(TIME):MAGNITUD100	0.765	0.778	0.249	0.660

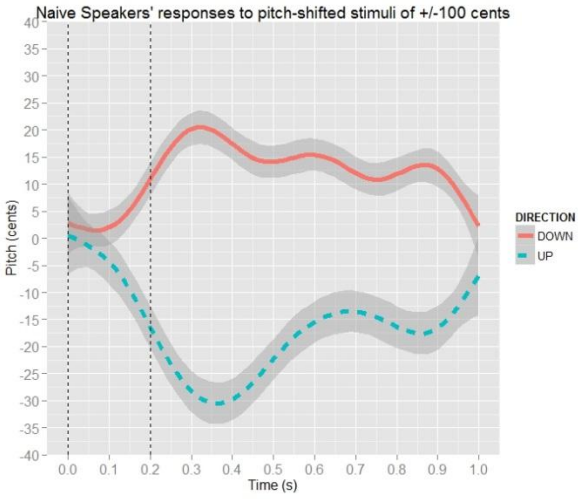
As the generalized additive models smoothed the F0 curves in temporal domain, the F0 contours resulting from the final model represent a smoothed curve in all cases. The estimated (smoothed) changes of the F0 contour in relation to the time dimension per condition built from the GAM are shown in Figure 3.6.

Figure 3.6 Estimated smoothed changes of the F0 contour in relation to the time dimension per condition built from the GAM. ***a.*** Naïve speakers' responses for +/-50 cents pitch-shift stimuli. ***b.*** Naïve speakers' responses for +/-100 cents pitch-shift stimuli. ***c.*** L2 learners' responses for +/-50 cents pitch-shift stimuli. ***d.*** L2 learners' responses for +/-100 cents pitch-shift stimuli. ***e.*** Mandarin speakers' responses for +/-50 cents pitch-shift stimuli. ***f.*** Mandarin speakers' responses for +/-100 cents pitch-shift stimuli.

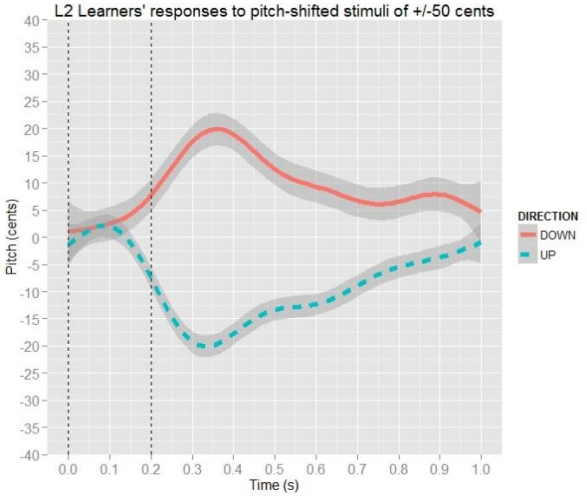
a.



b.



c.



d.

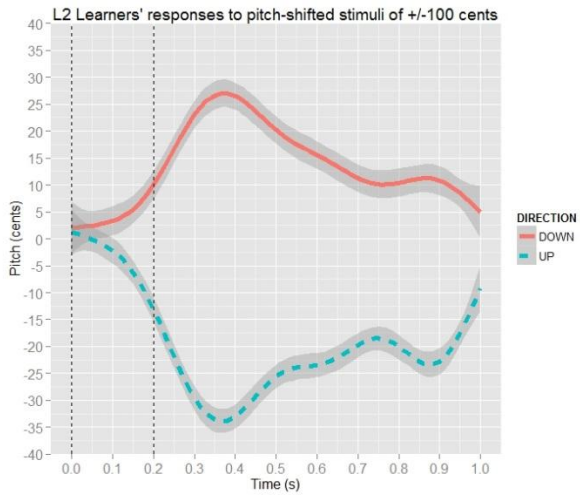
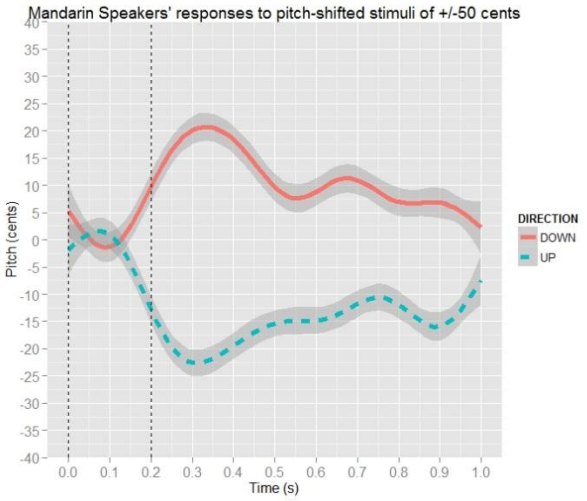
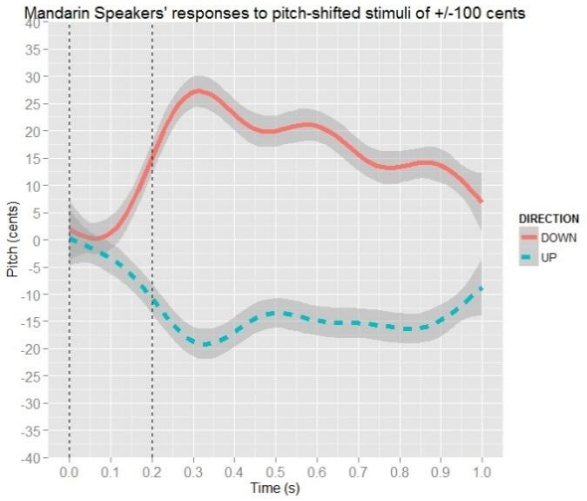


Figure 3.6 (Cont.)

e.



f.



Note. The x-axis represents the time in second. The onset of pitch-shift stimuli is at time 0 second and the offset is at time 0.2 second. The y-axis represents the amplitude of pitch-shift responses measured in cents. The red contours are compensatory responses to downward pitch perturbation. The blue contours are compensatory responses to upward pitch perturbation.

Note that the responses I included in the analysis are compensatory responses, so the direction of responses is opposite to the direction of pitch-shift stimuli. The onset of pitch-shift stimuli is at time 0 second and the offset is at time 0.2 second. Stratified post-hoc comparisons were also performed on the coefficients within each GROUP. Significance was evaluated with a Wald Chi-squared test for model coefficients. For naïve speakers, the responses to up-shift +100 cents stimuli were particularly larger than all the other conditions ($\chi^2(1)=8.8$, $p<0.01$). For L2 learners, the responses to pitch-shifted stimuli of +/-100 cents were overall larger than the responses to pitch-shifted stimuli of +/-50 cents (MAGNITUDE effect) ($\chi^2(1)=12.8$, $p<0.001$). There could be a second response (second peak) in the up-/down-shift 100 cents conditions. However, Mandarin speakers, unlike naïve speakers and L2 learners, were not affected by MAGNITUDE ($\chi^2(1)=1.5$, $p=0.22$).

3.4 Discussion

In this study, I investigated whether tonal language experience influences auditory discrimination and sensorimotor integration. In terms of discrimination, I found that both nonlinguistic tone discrimination and linguistic tone discrimination differentiate naïve speakers, L2 learners and Mandarin speakers. The results indicate that finer nonlinguistic tone discrimination (TD2) task is associated with better Mandarin tone discrimination (MD1), particularly when normalizing for tone variation across speakers. In terms of sensorimotor integration, even though the three groups did not show significantly different response amplitudes or response timing, they showed consistent differences when the whole F0 contour was compared. The F0 contour of the naïve speakers and L2 learners was affected by MAGNITUDE, while the F0 contour of Mandarin speakers was not affected by this factor. Using discriminant analysis, it was found that tone discrimination and pitch-shift responses are objective measurements that contribute to differentiating people with tonal language experience from those without tonal language experience. These promising but preliminary results suggest that lifelong experience speaking a tonal language may be accompanied by perceptual and sensorimotor differentiation from non-tonal speakers.

3.4.1 Perception and Production

The discrimination tests identified that L2 learners and Mandarin speakers showed more accurate tone discrimination than naïve speakers. Performance on MD2 also partially lead to separation of each groups, which is expected as this task loads on linguistic experience. The correlation analyses further showed that finer sensitivity on the nonlinguistic tone discrimination task (TD2) is associated with better discrimination on the first Mandarin tone discrimination task (MD1). An implication of these findings is that nonlinguistic tone discrimination ability may

predict an advantage for linguistic tone perception in certain L2 learners. Regarding the Mandarin group, the advantage shown in nonlinguistic tone discrimination is consistent with previous research that indicates the pitch characteristics of their first language can facilitate the ability to identify or produce a musical note (Deutsch, Henthorn, & Dolson, 2000; Henthorn & Deutsch, 2007). The findings follow these studies in suggesting that long-term exposure to a tonal language may benefit nonlinguistic tone discrimination. I also remain cautious about over-interpreting correlation results, but I believe a positive relationship between nonlinguistic tone discrimination and linguistic tone discrimination provides a compelling rational for further investigations of individual variation in tonal language aptitude.

The other important outcome of the perception tasks is that the Mandarin tone discrimination tasks, particularly MD2, successfully distinguished the naïve speakers from L2 learners while Mandarin speakers showed the best performance. The performance of the L2 learners on MD2 relative to naïve speakers suggests they are acquiring Mandarin tone perception abilities that can bridge across speaker differences. This is particularly indicated by the discriminant analysis result (with the use of Mandarin tone discrimination scores) that did not separate L2 learners from Mandarin speakers. In summary, naïve speakers may be able to use innate tone discrimination ability to distinguish tones of a single speaker, but it takes tonal language experience to consistently differentiate linguistic tonal variation across speakers. The L2 learners seem to be acquiring this ability which suggests their internal perceptual representations for Mandarin are gradually being modified.

As for production characteristics, here represented by pitch-shift responses, I did not find significant effects of language experience on the two data points (peak and onset) typically used in the pitch-shift paradigm. The experimental finding indicates that Mandarin speakers show basic pitch-shift amplitude and timing characteristics that approximate speakers of non-tonal

languages, when the stimulus is not a Mandarin word. Xu *et al.* (2004) reported that Mandarin speakers had larger pitch-shift amplitude and more rapid adjustments compared to their colleagues' work on native speakers of English (Burnett et al., 1998; Hain et al., 2000; Larson et al., 2001), but those statements were based on a comparison across different experiments of Mandarin speakers' production of tonal sequences with English speakers' production of sustained vowels. No comparison dataset was available to provide evidence of differences between language groups. In the current study, I administered the pitch-shift task to three language groups. The lack of significant differences in response peak amplitude and response timing could suggest that Mandarin speakers have larger pitch-shift amplitude and more rapid adjustment on voice F0 only when Mandarin speech tasks are used. Thus, testing for experience-dependent sensorimotor differences could require linguistically specific stimuli. Another potential reason for the lack of significant group differences in response peak amplitude and response timing could be the magnitude of pitch-shift stimuli (50 cents and 100 cents in the current study). Liu et al. (2010) and Chen et al. (2012) argue that Cantonese speakers had larger pitch-shift response amplitudes and larger neural responses than Mandarin speakers when the pitch perturbation was greater than 100 cents (200 and 500 cents). It is possible that Mandarin speakers only differ from naïve speakers and L2 learners with larger pitch perturbations. Study 2 will compare pitch shift responses to both Mandarin stimuli and simple vowel stimuli in naïve speakers, L2 learners and Mandarin speakers at larger pitch perturbations (200 cents) to isolate whether language experience alters pitch-response characteristics.

The novel follow-up analysis based on Kosling *et al.* (2013) and Sakai (2004) did suggest that pitch perturbations produced group differences in the whole F0 contours based on GAM analyses. Naïve speakers had larger F0 contours in the upward 100 cents condition. Mandarin speakers, however, showed no consistent effects of MAGNITUDE on their F0 contours,

suggesting that Mandarin speakers may maintain more stable vocalization patterns in the presence of unexpected changes in pitch feedback, which we will address further in the next section. As for L2 learners, whose F0 contours were affected by MAGNITUDE, the F0 contours of pitch-shift responses to +/-100 cent shifts were significantly larger than the F0 contours of pitch-shift responses to +/-50 cent shifts. L2 learners also differed from Mandarin speakers in showing overall larger F0 contours of responses to pitch perturbations. This suggests that L2 learners have not acquired a native-like/stable vocal control and thus are more likely to be affected by the pitch perturbation.

The discriminant analysis generated sizeable factor coefficients that suggest pitch-shift responses and discrimination contribute to speaker classification. A linear combination of pitch-shift and tone discrimination characteristics was required to separate Mandarin speakers from L2 learners. The discriminant analysis and F0 contour comparison suggest that internal phonatory control in Mandarin speakers differs from L2 learners and naïve speakers. In particular, the L2 learners have not acquired sensorimotor responses that mirror Mandarin speakers. Further learning may still be required among the L2 learner group to acquire the linguistically specific motor control found in native speakers of tonal languages.

3.4.2 Internal Models for Language Learning

As per the introduction, I suggest that internal model theory is relevant for the understanding of linguistic aspects of speech production. It means that the brain 'internalizes' representation of new and established motor skills. Related to language acquisition, the native speaker's brain is continuously modeling the relationships between the desired speech movement and its expected outcome to learn and eventually master the ambient language. If so, then a native speaker's response to auditory perturbation should reflect their language specific internal

model. The current finding that F0 contours in Mandarin speakers were less affected by MAGNITUDE bears a certain similarity to the suppression of pitch perturbations reported by Liu, Xu, et al. (2009). So instead of a heightened sensitivity, tonal language learning may contribute to vocal control stability. Given this pattern, the sensorimotor goal of tonal language learning in L2 learners may involve learning to produce stable F0 contours across varied speaking and auditory feedback conditions. Rapid correction of responses may be less important than consistent and stable vocal control in their attempts to produce Mandarin tones. L2 learners may still be in the stage of reshaping their internal models for tone perception/production to be more native-like. A possible indicator of the reshaping process may be seen in that L2 learners were sensitive to the magnitude of pitch perturbation. I predict that L2 learners will resemble Mandarin speakers in their ability to suppress pitch perturbation when their vocal control of lexical tone is stabilized.

3.5 Conclusion

I have taken the first step in examining both perception and audio-vocal responses in the context of tone learning for three different language groups, including naïve speakers, L2 learners, and Mandarin speakers. Clear advantages were found for nonlinguistic and linguistic tone discrimination in Mandarin speakers and L2 learners compared to naïve speakers. The F0 contours that may represent how internal models of vocal control stabilize the voice were significantly different between groups. This suggests that Mandarin speakers may have more stable internal models. The results also show that both tone perception and pitch-shift responses are objective measurements that can effectively classify the individuals in three language groups. The research adds to the current knowledge gap on whether vocal control mechanisms in L2 learners can be remodeled by tonal language learning. This new information contributes to a

more unified understanding of language learning and the underlying changes in the language system. It opens the door to further studies of potentially malleable neural mechanisms involved in tonal language acquisition.

CHAPTER 4
STUDY 2: TONE PERCEPTION AND SENSORIMOTOR RESPONSES TO MANDARIN SYLLABLES

4.1 Introduction

Study 2 is the most critical study that is documented in this dissertation. It brings together the critical issues that were not addressed in Study 1 and addresses the limitations in previous literature. First, actual Mandarin stimuli were compared against simple vowels. Study 1 only examined pitch-shift responses using simple vowels. To understand language-dependent and stimulus-specific modulation of pitch-shift perturbation, it requires comparisons of Mandarin stimuli and non-native stimuli in speakers with different language experiences. Study 2 presented in this chapter will use tone stimuli to investigate whether the suppressed responses are due to more stable internal models of native speakers of tone languages. Second, appropriate groups were selected: naïve speakers, trained vocalists, L2 learners, native speakers of Mandarin. In Study 1, musical training experience was not controlled, but it could be a factor facilitating the learning of new audio-vocal interactions. Singing experience (or aptitude) could presumably be related to superior tone discrimination and tone production abilities. Study 2 will examine trained singers in addition to naïve speakers, L2 learners and Mandarin speakers and investigate how tone perception and audio-vocal responses vary among the four groups. Third, the relationship between Mandarin tone production and Mandarin tone perception was evaluated, by using an adaptive Mandarin tone discrimination task. The adaptive program captures tone learning processes and should provide an informative picture of how stepwise progression can differentiate the four language groups.

In terms of voice production, experienced singers appear to have enhanced control over voice F0 than non-musicians because they show the capability of suppressing pitch-shift

responses (Jones & Keough, 2008; Zarate & Zatorre, 2008). An attenuation effect suggests experienced singers may be able to block out the potentially perturbing feedback to remain on key. Attenuation effects in both experienced singers (Jones & Keough, 2008; Zarate & Zatorre, 2008) and Mandarin speakers (Liu, Xu, et al., 2009) raise the possibility that experience controlling vocal tone has the added potential of being able to 'ignore' or 'cancel-out' potentially disruptive sensory feedback. To extend the relevance of this study, I will also examine whether experienced singers can extend their attenuation to linguistically meaningful stimuli such as Mandarin words, i.e., whether experienced singers behave like Mandarin speakers regardless of the vocalization task.

In the current study, I draw together three important factors that may contribute to differences in L2 learning of Mandarin: i) language experience, ii) vocal training experience, and iii) stimulus-specificity (i.e., different Mandarin tones). Regarding language experience, native speakers of Mandarin are expected to have robust internal models of tone production, while non-tonal speakers (naïve speakers) are not. L2 learners of Mandarin should be in process of acquiring the internal models. Regarding vocal training experience, because trained vocalists should have better abilities to discriminate tone and to control voice F0 control, their internal models for tone may potentially be similar to Mandarin speakers' internal models. Regarding stimulus-specificity, it is possible that speakers' internal models of voice production may regulate musical pitch and linguistic pitch differentially and regulate different Mandarin tones differently, with the functional specificity being dependent on language and vocal training experiences. In this study, I compared Mandarin bi-tonal sequences and the sustained vowel /a/ in the pitch-shift paradigm to examine pitch-shift responses for four groups, including naïve speakers without exposure to tonal languages, trained vocalists, L2 learners of Mandarin and native Mandarin speakers.

I asked the following questions: i) whether trained vocalists and L2 learners of Mandarin bear a resemblance to Mandarin speakers in terms of both pitch-shift responses and Mandarin tone perception, ii) whether pitch-shift responses are specific to linguistic or non-linguistic stimuli across the four groups, iii) whether tone learning processes in the adaptive Mandarin tone discrimination task can differentiate the four groups, and iv) whether tone learning processes in the Mandarin tone discrimination task is related to pitch-shift responses among the four groups. My hypotheses are that i) Mandarin language or vocal training experience will be associated with greater attenuation of pitch-shift responses in general and enhanced perception of tonal contrasts, ii) pitch-shift responses will be stimulus-specific in that Mandarin speakers will have greater attenuation of pitch-shift responses during Mandarin tone production, iii) there should be an advantage in Mandarin tone learning for L2 learners and trained vocalists, but not for naïve speakers, and iv) tone learning processes in the Mandarin tone discrimination task should be related to the suppression of pitch-shift responses among the four groups.

4.2 Experiments in Study 2

The three tasks conducted in Study 2 are presented below. The nonlinguistic tone discrimination task examines the participants' ability to discriminate two tones with different fundamental frequency. The adaptive Mandarin tone discrimination task explores the participants' ability to differentiate lexical tone differences and examines if the participants can get to the most difficult level at the end of the adaptive program. The pitch-shift task investigates whether sensorimotor control over voice F0 is dependent on language experience or vocal training experience and is subject to stimulus type. The same participants participated in all the three tasks.

4.2.1 Participants

The thirty-six participants in the experiment included: Ten naïve speakers (7 male) who were never exposed to tonal languages and whose native language is American English, 6 trained vocalists (3 male) whose major is voice and whose native language is American English, 10 adult L2 learners of Mandarin (7 male) whose native language is either American English or Korean, and 10 native speakers of Mandarin (4 male). All the nonmusician groups (naïve speakers, L2 learners, and native speakers of Mandarin) included some participants with musical training before college, but none possessed vocal training or had ongoing participation in formal musical training. Based on the post-hoc interviews, the average period of musical training (instruments) before college was 5.4 ± 1.4 years for naïve speakers, 13.8 ± 2.3 years for trained vocalists, 7.7 ± 2.1 years for L2 learners, and 4.5 ± 1.8 years for Mandarin speakers. All were recruited from University of Illinois at Urbana-Champaign. All participants passed a hearing screening at 20 dB HL bilaterally at 125, 250, 500, 750, 1000, 2000, 3000, and 4000 Hz. None of the participants reported a history of neurological, communication, hearing or language disorders. They were paid $10 for their participation. All methods reported herein were approved by the Institutional Review Board at the University of Illinois and all participants provided written informed consent. The entire experiment took place in a single session lasting about 60 minutes.

Participants were seated in a sound booth and wore Sennheiser HD-280 Pro headphones and a headworn Shure microphone placed 2 cm away from the corner of the mouth. They were asked to complete the three tasks in the following order: nonlinguistic tone discrimination task, adaptive Mandarin tone discrimination task and pitch-shift task.

4.2.2 Nonlinguistic Tone Discrimination Task

4.2.2.1 Materials

The nonlinguistic tone discrimination task examines the participants' ability to differentiate the pairs of tones that differ in fundamental frequency. To estimate a listener's ability for nonlinguistic tone discrimination, the Adaptive Pitch Test (Mandell, 2009) available on the tonometric website was presented through headphones via a laptop to each participant.

4.2.2.2 Procedures

The procedure was the same as in Study 1 (Chapter 3), except that no accuracy feedback was provided in this study. Accuracy feedback (correct or wrong for each trial) was covered up so the participant could not track their performance. After determining the smallest tone difference in Hertz that the participant can consistently discriminate, the test terminates and the just-noticeable difference (JND) in Hertz (the threshold at which a change is perceived (Kollmeier, Brand, & Meyer, 2008)) was presented on the webpage. Participants repeated this task twice (i.e., 2 attempts) to obtain a reliable estimate of pitch discrimination. Each attempt took about 3 minutes to complete.

4.2.3 Adaptive Mandarin Tone Discrimination Task

4.2.3.1 Materials

The adaptive Mandarin tone discrimination task was conducted to investigate whether participants can get to the most difficult level and if the patterns of stepwise progression are dependent on language experience or vocal training experience.

The adaptive test adopted the monosyllabic Mandarin tone stimuli used in Shih et al. (2010). Shih and colleagues found that adaptive tone identification training was most beneficial for adult

L2 Mandarin learners. The stimuli were monosyllabic Mandarin words recorded at different talker-to-listener distances (TLD (Cheyne et al., 2009; Pelegrin-Garcia et al., 2011)). The tone identification results from 36 Mandarin learners showed a non-linear response to speech recorded at different TLD. Speech recorded at a close distance was difficult for L2 learners due to reduction, while those recorded at a long distance was harder due to exaggeration. Speech files recorded with a TLD distance of 8 feet (step 4 of Shih's *et al*.) had the highest identification rate by L2 learners. L2 Learner's tone identification accuracy decreased linearly from step 4 (TLD of 8 feet) to step 11 (TLD of 20 feet), when an aberrant bump at step 10 was excluded. Following this result, in the current study, the speech files that varied in identification difficulty based on TLD were employed within a 7-staircase adaptive test to determine how Mandarin tone discrimination varies among the four groups. The design of the seven stairs (1: easy ~ 7: difficulty) is presented in (1). In each stair, there were 28 tokens in the stimuli bank with 18 pairs of different tones and 10 pairs of the same tone. Among the 28 tokens, there were 13 pairs of the most confusing tone2-tone3 combination (C.-Y. Lee, Tao, & Bond, 2010; Shih et al., 2010). The two female voices and one male voice had an approximately equal number of occurrences.

(1) Stair 1: Shih *et al.*'s step 4; same segments and same speaker's voice

Stair 2: Shih *et al.*'s step 7; same segments and same speaker's voice

Stair 3: Shih *et al.*'s step 9; same segments and same speaker's voice

Stair 4: Shih *et al.*'s step 11; same segments and same speaker's voice

Stair 5: Shih *et al.*'s step 4; different segments and same speaker's voice

Stair 6: Shih *et al.*'s step 4; same segments and different speakers' voices

Stair 7: Shih *et al.*'s step 4; different segments and different speakers' voices

4.2.3.2 Procedures

E-Prime software (v.1, Psychology Software Tools, Inc., Sharpsburg, PA) was utilized to present pairs of Mandarin tones that participants were required to discriminate. Participants listened to pairs of Mandarin words that varied in tone pairs or in segments (consonants and vowels). Participants were asked to judge whether the two words had the same or different tonal contours by pressing a "same" or "different" button. Participants had only one chance to listen to the pair.

Participants started at stair 4. The stepping rule follows the algorithm in Shih *et al.*'s tone training program: If they correctly identify 12 stimuli in a row, they advance to a more difficult level; on the other hand, if they make two mistakes in a row, they drop to an easier level. This excludes the possibility that participants attain better performance by guessing. The adaptive test terminated after 52 trials. The stair where participants end represents their proficiency level of Mandarin tone discrimination. Participants were given a short practice session (12 trials) and their overall accuracy was presented on the screen at the end.

4.2.4 Pitch-shift Task

4.2.4.1 Materials

The pitch-shift task examined whether the suppression of the pitch-shift responses depends on language/musical experience or stimulus type. The materials are presented in (2).

(2) Pitch-shift stimuli in Study 2

a. /ma1 ma1/

b. /ma1 ma2/

c. /ma1 ma4/

d. /a/

In the Mandarin tone condition, participants had to produce /ma ma/ sound which varied in tonal contour on the second syllable (/ma1 ma1/ (T11 for short), /ma1 ma2/ (T12 for short) and /ma1 ma4/ (T14 for short)). Each syllable pair was approximately 1.5 second in duration. At 800 ms after vocalization onset (which corresponded to the beginning of the second /ma/), the pitch of the feedback signal was altered for 200 ms by a ±200 cent shift with shift direction randomized. Two vocalization sessions were recorded before the formal pitch-shifting was implemented. The first session included 15 trials (5 repetitions for each bi-tonal sequence) without pitch-shifted feedback so baseline performance on tone production could be recorded. Then a second session of 15 trials was used to introduce pitch-shifted feedback to the participants. In the formal pitch-shifting session, 70 trials (30 repetitions of T11, 20 repetitions of T12 and 20 repetitions of T14) were recorded with one pitch-shift perturbation per trial with the number of upward and downward shifts approximately equal.

In the sustained vowel condition, participants had to produce the simple vowel /a/. At a randomized interval (200, 300, and 400 ms) after vocalization onset, the pitch of the feedback signal was altered for 200 ms by a ±200 cent shift with shift direction randomized. Thirty vocalizations were recorded with two pitch-shift perturbations per trial for a total of 60 perturbations with the number of upward and downward shifts approximately equal.

4.2.4.2 Procedures

Participants first heard a female /ma ma/ sound recording (T11, T12, or T14), or a male /a/ sound recording. Participants were trained to produce Mandarin bi-tonal sequences /ma1 ma1/, /ma1 ma2/, /ma1 ma4/ and the vowel /a/ at approximately 70 dB sound pressure level (SPL) while self-monitoring their vocal volume on a Dorrough Loudness Monitor (model 40-A) placed in front of them. The voice signal from the microphone was amplified with a YAMAHA mixer (MG102c) and sent to an Eventide Ultra-Harmonizer (model H7600) that generated the pitch shifts. The participant's own voice signal was amplified with a Samson S-phone Headphone Amplifier and played back to him/her at approximately 80 dB SPL to reduce the possible influence of bone conduction. MIDI software (MAX/MSP v.5 by Cycling 74, Walnut, CA) connected to a MOTU (model UltraLite-mk3 Hybrid) controlled the timing, duration and magnitude of the pitch shifts produced by the Eventide ultra-harmonizer. The participant's voice, altered feedback signal, and pitch-shift events (transistor transistor logic (TTL) pulses) were digitized at 5 kHz per channel with WINDAQ/Pro software (DATAQ Instruments, Inc., Akron, OH).

In the Mandarin tone condition, participants were instructed to repeat or imitate the tonal contours as well as the production length and to ignore any changes in what they heard. In the sustained vowel condition, participants were instructed to hold vocal pitch and volume as steady as possible for 5 seconds and to ignore any changes in what they heard. The stimuli (Mandarin bi-tonal sequences with tone indication and the vowel /a/) were presented to the participant on a Dell laptop running E-prime software (v.1, Psychology Software Tools, Inc., Sharpsburg, PA).

4.2.5 Data Analysis

In the pitch-shift task, the raw signals in WINDAQ were imported into MATLAB (R2009b,

The Mathworks Inc., Natick, MA). In the Mandarin condition, the signals were then sorted into files based the tonal category and the direction of the pitch-shift stimuli (up-shift vs. down-shift). Vocalizations where the shift did not occur at the beginning of the second /ma/ were excluded. For each vocalization, the onsets of the first /ma/ and the second /ma/ and the offset of the second /ma/ were marked. The *pitch values* were tracked using the SWIPE pitch estimation algorithm (10 ms interval) (Camacho & Harris, 2008; Scheerer & Jones, 2012) and then converted into cents (cents = $1200*\log_2(F0/\text{baseline F0})$), where the baseline F0 is the mean F0 of the first syllable /ma/. In order to average across trials and speakers, the cents records across both syllables were time normalized to 2 s by using the linear interpolation function in MATLAB (Xu et al., 2004). Following responses were excluded in this study so that only compensatory responses are considered. Then the *mean response amplitude difference* between up-shift and down-shift stimuli was obtained for the first /ma/ (*Syl1_MeanDiff*), by taking the difference between up-shift and down-shift at every time point and then computing the average of the data points. The *mean response amplitude difference* (*Syl1_MeanDiff*) is expected to be small as no pitch-shift occurred on the first /ma/. The *mean response amplitude difference* and *maximum response amplitude difference* between up-shift and down-shift stimuli were obtained for the second /ma/ where pitch-shift occurred (*Syl2_MeanDiff* and *Syl2_MaxDiff*), by taking the difference between up-shift and down-shift at every time point and then finding the average and maximum differences among the data points. Small values of *Syl2_MeanDiff* and *Syl2_MaxDiff* are expected for cases where pitch-shift responses are attenuated (unaffected by the shifted direction).

In the sustained vowel condition, the signals were sorted into files based on the direction of pitch shift stimuli (up-shift vs. down-shift). For each vocalization, a 1.2 second window including a 200 ms pre-pulse baseline, the 200 ms pulse, and an 800 ms post-pulse period was selected for acoustic analysis. The *pitch values* were tracked using the SWIPE pitch estimation

algorithm (10 ms interval) (Camacho & Harris, 2008) and then converted into cents (cents = $1200*\log_2(F0/\text{baseline } F0)$), where the baseline is the mean F0 between 0-200 ms. Following responses and responses without a steady baseline (the phase before pitch-shift occurs) were excluded so that only compensatory responses are considered. Then the onset time, peak time, and peak amplitude of the average responses were obtained in the following way: *onset time* was defined as the time at which the response exceeded 2 standard deviations of the pre-stimulus mean; *peak time* and *peak amplitude* corresponded to the first peak of averaged response that usually occurred between 200-600 ms after the onset of the pitch-shift stimulus. Each of these points was selected with an automated peak-picking algorithm in MATLAB.

To evaluate nonlinguistic tone discrimination, a permutation test with repeated measures ANOVA fitting the linear model of GROUP x ATTEMPT was performed. To evaluate adaptive Mandarin tone discrimination, a logistic regression of accuracy was conducted to examine the main effect of GROUP. To investigate the advancement (growth) in the adaptive Mandarin tone discrimination, hierarchical linear models (also called multilevel models) were used to analyze repeated observations from an individual ordered by time/trial number (level-1) and treat the variable GROUP (individual characteristics) as a level-2 predictor. Correlations were conducted to examine whether nonlinguistic tone discrimination performance was related to Mandarin tone discrimination accuracy.

To evaluate pitch-shift responses, a one-way ANOVA comparing the GROUP effect on each dependent variable (*Syl1_MeanDiff*, *Syl2_MeanDiff* and *Syl2_MaxDiff*) was conducted separately for T11, T12, T14; in addition, a repeated measures ANOVA fitting the linear model of GROUP x DIRECTION was performed separately on each dependent variable *onset time*, *peak time*, and *peak amplitude* for /a/. Correlations were conducted to examine whether the pitch-shift responses to /a/ and the pitch-shift responses to Mandarin bi-tonal sequences were associated and whether

pitch-shift responses were associated to perception performance.

4.3 Results

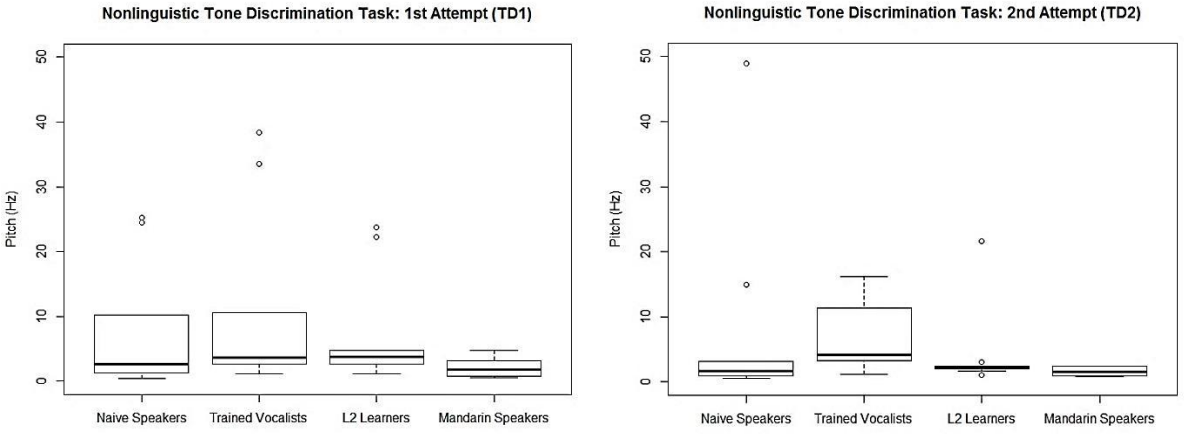
This section presents the results of nonlinguistic tone discrimination, adaptive Mandarin tone discrimination task and pitch-shift task. Since the perception data and production data violated the assumptions of normal distribution and homogeneity of variance, permutation tests were used to examine statistical significance.

4.3.1 Nonlinguistic Tone Discrimination Task

The scores on the nonlinguistic tone discrimination task (TD) are depicted in Figure 4.1. In the nonlinguistic tone discrimination task, a lower score in Hz corresponds to better discrimination ability. The permutation tests with repeated measures ANOVA show there were no significant main effects of GROUP ($F(3,64) = 1.992$, $p = .119$) or ATTEMPT ($F(1,64) = 1.151$, $p = .304$), and no significant interaction ($F(3,64) = 0.523$, $p = .881$) on nonlinguistic discrimination scores.

Figure 4.1 ***a.*** Discrimination performance in the first attempt (TD1) of nonlinguistic tone discrimination task. ***b.*** Discrimination performance in the second attempt (TD2) of nonlinguistic tone discrimination task.

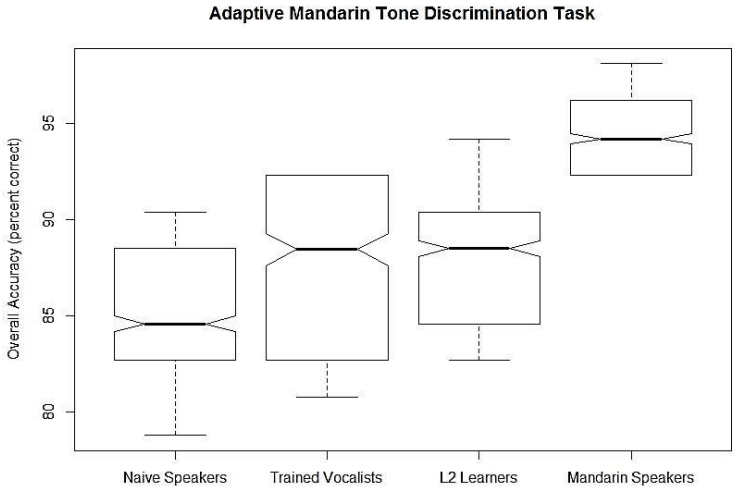
a. b.



4.3.2 Adaptive Mandarin Tone Discrimination Task

The overall accuracy in the adaptive Mandarin tone discrimination task (MD) is depicted in the boxplots in Figure 4.2. For the adaptive Mandarin tone discrimination task, a higher score (percent correct) corresponds to better discrimination ability. A binary logistic regression was conducted to regress the accuracy of each trial (1 for correct and 0 for incorrect) on GROUP. The model shows that there was a significant effect of GROUP (*Wald* $\chi^2(3) = 23.45$, $p < .0001$). Pairwise comparisons with the Tukey's Honest Significant Difference method show that the Mandarin group performed significantly better than all the other three groups ($p < .01$), whereas no significant differences were found for any other pairs (**MANDARIN>L2 LEARNER=VOCALIST=NAÏVE**).

Figure 4.2 Overall accuracy in the adaptive Mandarin tone discrimination task by group (MD).



To model the advancement (growth) in the adaptive Mandarin tone discrimination, hierarchical linear models were used to fit repeated observations from an individual ordered by time/trial number (level-1) and the GROUP variable (level-2). The goal in growth modeling is to determine whether there were consistent patterns in the relationship between time (trial number) and stair value, i.e., whether participants can all advance to the highest level in the adaptive test. I follow the model comparison approach outlined in Bliese and Ployhart (2002).

For the level-1 analyses, I first estimated the null model and calculated the Intraclass Correlation Coefficient (ICC) to determine whether the stair value randomly varied among individuals. The ICC associated with the stair value is .30, which indicates that only 30% of the variance in any individual stair value can be explained by the properties of the individual. This suggests that individuals tend to increase or decrease stair value over time/trial number and that the growth varied among individuals. The individual plot is shown in Figure 4.3. For example, the naïve speaker labeled as 01 stayed in stair 4 (the beginning level) for 21 trials, and then advanced to stair 5 in the 22nd trial and to stair 6 in the 34th trial. He ended up with stair 6 when

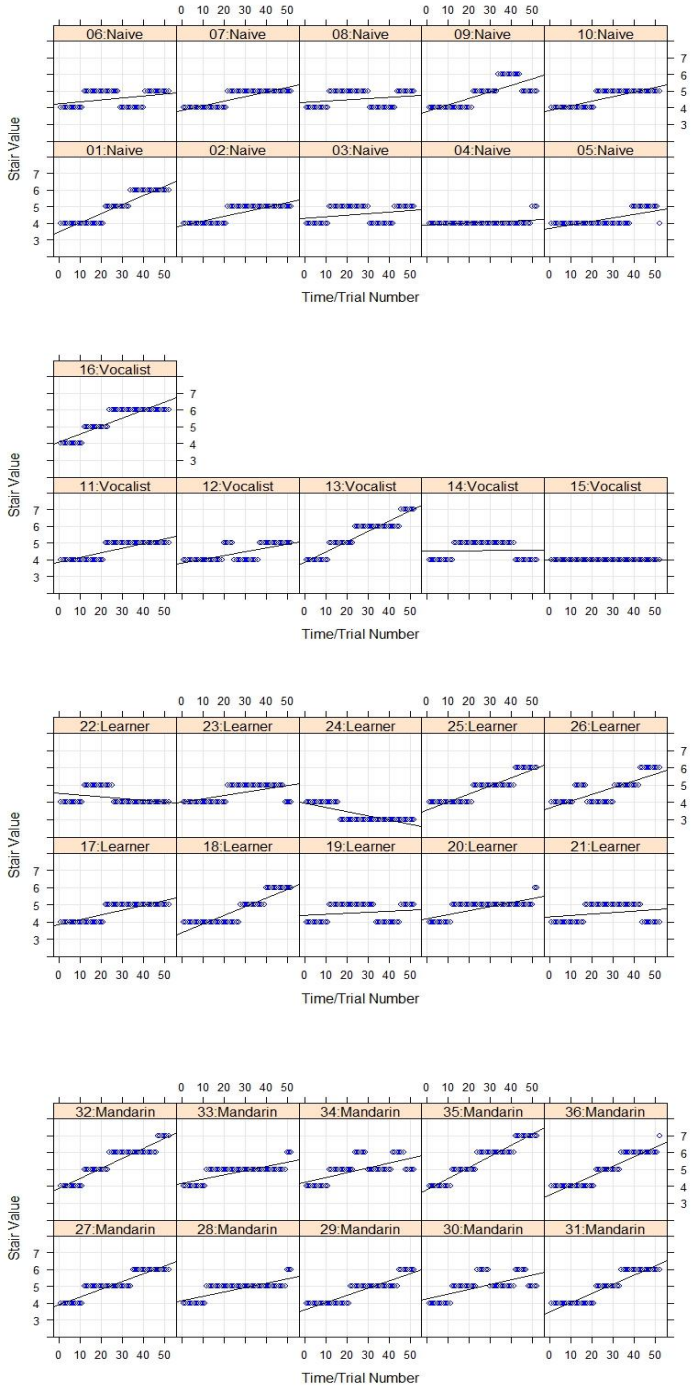
the adaptive program ceased.

Second, I examined the relationship between time/trial number and stair value (i.e., whether the stair value generally increases, decreases, or shows some other type of change over time) by regressing stair value on time/trial number (fixed effect) in a model with a random intercept. The results show that there was a significant linear relationship ($t(1834) = 15.219$, $p < .0001$) and a significant quadratic trend ($t(1834) = -6.110$, $p < .0001$) between time/trial number and stair value.

Third, I investigated whether the linear and quadratic relationships between time/trial number and stair value was constant among individuals or varies on an individual-by-individual basis. This was done by fitting a random slope model with an addition of the linear effect for time/trial number as a random effect. The results show that the random slope between time/trial number and stair value fit the data better than a model that fixed the slope to a constant value for all individuals (Log-likelihood ratio = 952.561, $p < .0001$).

Fourth, as the errors (residuals) could be correlated for longitudinal data, I included an autocorrelation structure for error to account for dependencies over time. When a first order autoregressive structure was used in the covariance matrix for the errors, the model fit the data better than a model that assumed no autocorrelation (Log-likelihood ratio = 2294.347, $p < .0001$). In sum, the level-1 analyses show that i) the individuals tended to increase or decrease stair value over time/trial number in a linear or a quadratic fashion, 2) the linear and quadratic relationships between time/trial number and stair value varied among individuals, and 3) there was a significant autocorrelation for the errors in the data.

Figure 4.3 The advancement (growth) in the adaptive Mandarin tone discrimination task (MD) by individual.

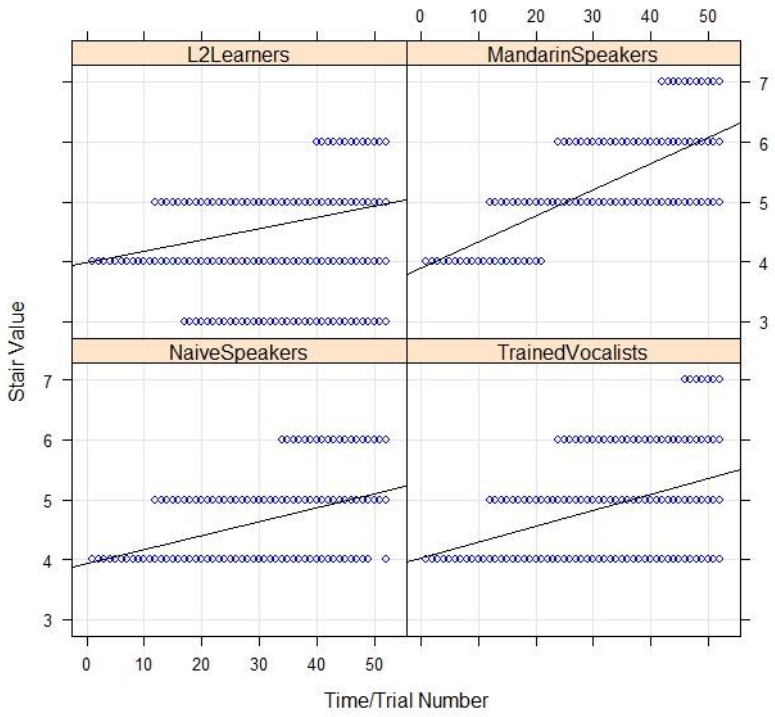


Note. The x-axis represents time/trial number, i.e., from the 1st trial to the 52nd trial. The y-axis represents stair value (see section 4.2.3.1 on p.82). The first panel indicates the results from 10 naïve speakers, the second panel from 6 trained vocalists, the third panel from 10 L2 learners, and the fourth panel from 10 native speakers of Mandarin. The straight lines indicate the linear

trends between time/trial number and stair value.

Next, I added level-2 predictors of intercept and slope variance, in order to examine whether the factor GROUP can explain why some individuals had advancement while other individuals did not. First, as all individuals started at stair 4, the intercept of stair value should not vary with the GROUP. The inclusion of GROUP as a new fixed effect shows that all four groups of participants did not different significantly from the starting stair value ($t(1835) = 0.00002$, $p = .999$). Second, consider the slope variation. I attempted to determine whether participants' groups can explain the variation in the time-stair value slope. This was done by including an interaction term for time/trial number and GROUP. When the interaction term was added, the quadratic trend between time/trial number and stair value became insignificant ($t(1828) = -0.391$, $p = .696$). Thus, I dropped the quadratic term and refit the data by using the linear trend between time/trial number and stair value. The results show that the slope for the Mandarin group was significantly different from other groups, which means Mandarin speakers advanced (over time) more than the others ($p < .01$) (see Figure 4.4).

Figure 4.4 The advancement (growth) in the adaptive Mandarin tone discrimination task (MD) by group.



Note. The x-axis represents time/trial number, i.e., from the 1st trial to the 52nd trial. The y-axis represents stair value (see section 4.2.3.1 on p.82). The straight lines indicate the linear trends between time/trial number and stair value.

4.3.3 Correlation between Two Perception Tasks

To further investigate whether performance in the nonlinguistic tone discrimination task was correlated with the Mandarin tone discrimination task, distribution-free spearman correlation analyses between the TD scores and the adaptive MD overall accuracy were conducted. Neither TD1 (the first attempt, r = .160, p = .352) nor TD2 (the second attempt, r = .235, p = .168) were significantly correlated with MD. Similarly, within each group, neither TD1 nor TD2 were significantly correlated with MD.

4.3.4 Post-hoc Analysis

The post-hoc interviews with all participants on the background of musical training show that participants with a longer period of musical training (such as piano, violin, flute, etc.) had better nonlinguistic tone discrimination (TD1: $r = -0.363$, $p < .05$; TD2: $r = -0.504$, $p < .01$). The period of musical training was only associated with nonlinguistic tone discrimination, not associated with Mandarin tone discrimination (MD: $r = .067$, $p = .854$).

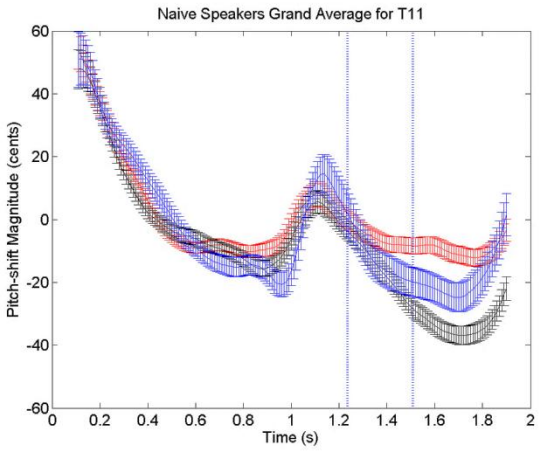
4.3.5 Pitch-shift Task

4.3.5.1 Mandarin Bi-tonal Sequences: T11, T12, and T14

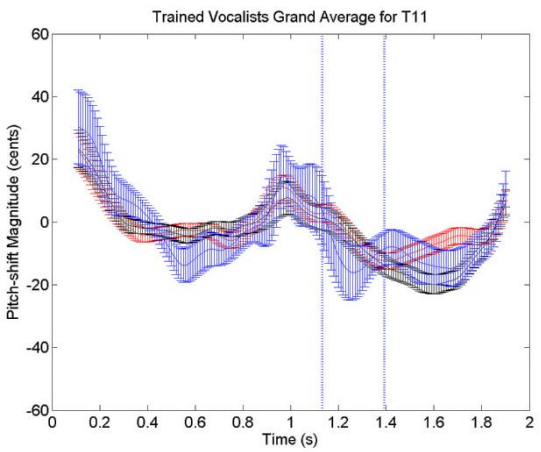
In the Mandarin condition, I conducted a separate permutation test with ANOVA fitting the factor of GROUP for T11, T12, and T14 on each dependent variable (*Syl1_MeanDiff*, *Syl2_MeanDiff* and *Syl2_MaxDiff*), to investigate whether pitch-shift responses are affected by language experience or vocal training experience. *Syl1_MeanDiff* indicates the *absolute mean response amplitude difference* between up-shift and down-shift F0 traces for the first /ma/. *Syl2_MeanDiff* and *Syl2_MaxDiff* indicate the *absolute mean* and *absolute maximum response amplitude differences* between up-shift and down-shift stimuli obtained for F0 traces of the second /ma/, respectively, where the pitch-shift compensation occurred. Small values of *Syl2_MeanDiff* and *Syl2_MaxDiff* are expected for cases where pitch-shift responses are attenuated (unaffected by the shift direction). The pitch-shift responses are displayed in Figure 4.5-Figure 4.7.

Figure 4.5: Pitch-shift responses during T11 production by group (zoomed in to ±60 cents).

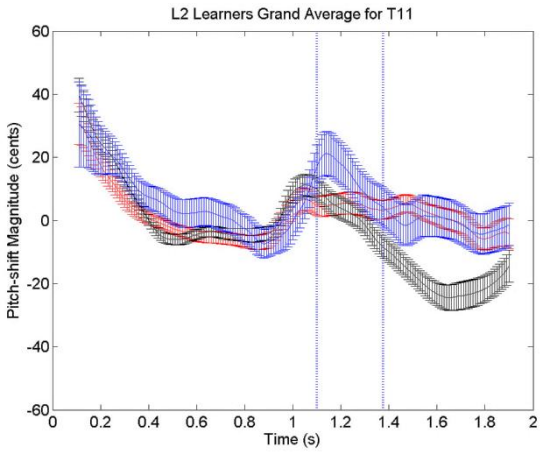
a.



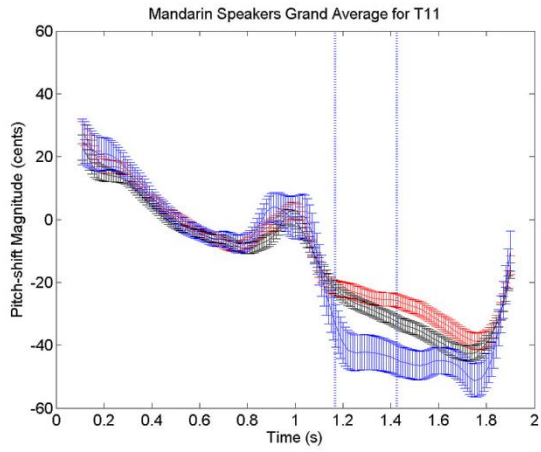
b.



c.



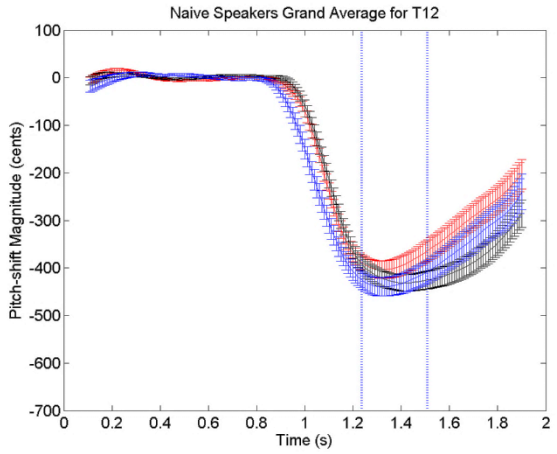
d.



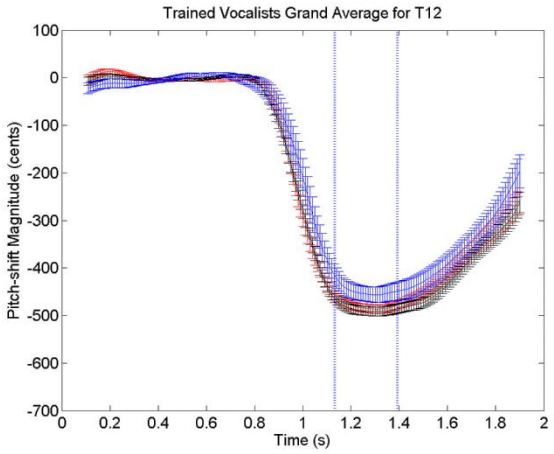
Note. The time on the x-axis is normalized time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Black curves represent the responses to up-shift stimuli. Blue curves represent the responses to the controls (no shift). The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

Figure 4.6 Pitch-shift responses during T12 production by group.

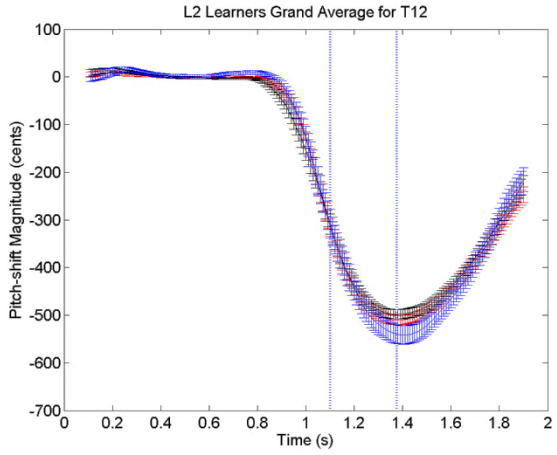
a.



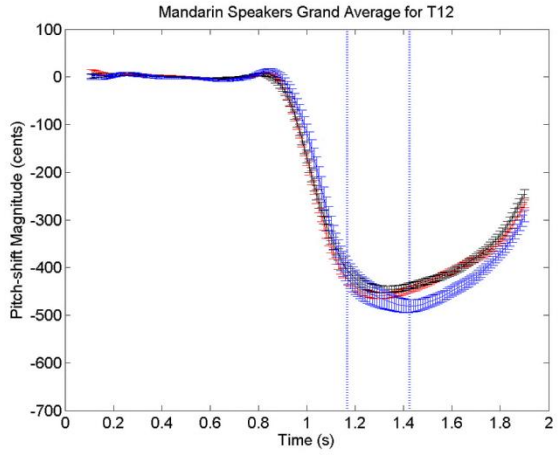
b.



c.



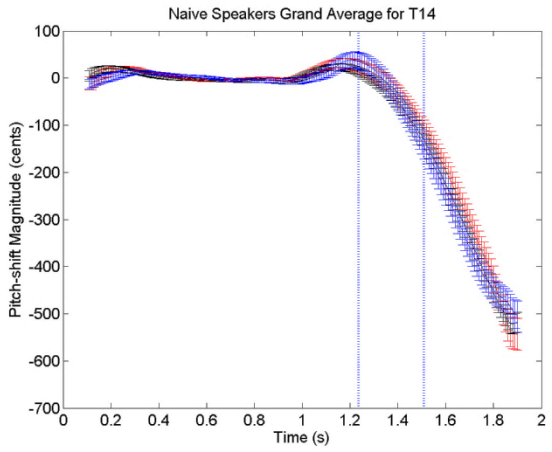
d.



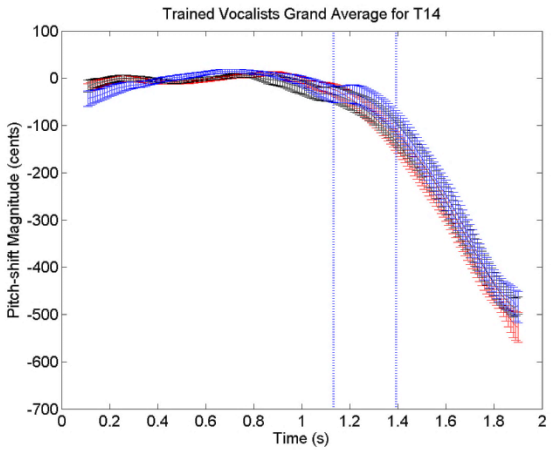
Note. The time on the x-axis is normalized time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Black curves represent the responses to up-shift stimuli. Blue curves represent the responses to the controls (no shift). The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

Figure 4.7 Pitch-shift responses during T14 production by group.

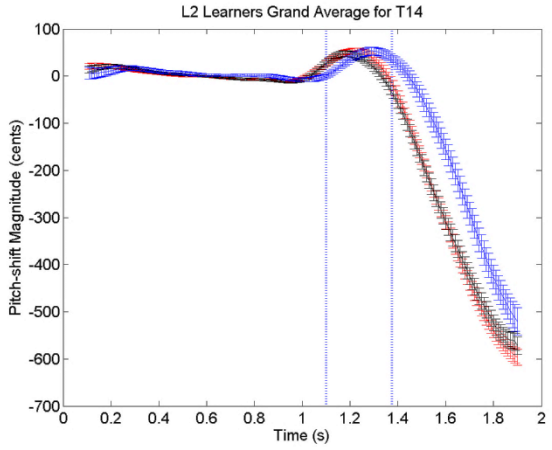
a.



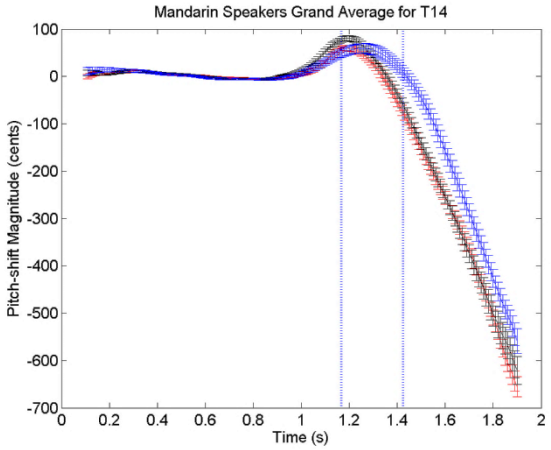
b.



c.



d.



Note. The time on the x-axis is normalized time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Black curves represent the responses to up-shift stimuli. Blue curves represent the responses to the controls (no shift). The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

For the T11 comparison, there was no significant GROUP effect on the *Syl1_MeanDiff* ($F(3,32) = 1.841$, $p = .201$), showing that the vocalization was stable in the baseline (the first /ma/). However, there was a significant GROUP effect on the *Syl2_MeanDiff* ($F(3,32) = 2.724$, $p < .05$). There was also a significant GROUP effect on the *Syl2_MaxDiff* ($F(3,32) = 3.277$, $p < .05$).

Post-hoc comparisons on *Syl2_MeanDiff* with the Least Significance Difference method show that the L2 learners had significantly larger mean magnitudes between up-shift and down-shift on the second /ma/ than the Mandarin group (diff = 3.387, $p < .05$) and the trained vocalists(diff = 3.478, $p < .05$). Post-hoc comparisons with the Tukey's Honest Significant Difference method and the Least Significance Difference method on *Syl2_MaxDiff* show that the naïve group had significantly larger magnitude deviations between up-shift and down-shift stimuli on the second /ma/ than the Mandarin group (diff = 20.662, $p < .05$) but no other group differences were detected.

For the T12 comparison, there was no significant GROUP effect on the *Syl1_MeanDiff* ($F(3,32) = 2.531$, $p = .07$). There was no significant GROUP effect on the *Syl2_MeanDiff* ($F(3,32) = 1.952$, $p = .127$). However, a significant GROUP effect was found on the *Syl2_MaxDiff* ($F(3,32) = 3.361$, $p < .05$). Post-hoc comparisons with the Tukey's Honest Significant Difference method show that the naïve group had nearly significantly larger magnitude deviation between up-shift and down-shift on the second /ma/ (*Syl2_MaxDiff*) than the Mandarin group (diff = 31.542, p=.054). Post-hoc comparisons with the Least Significance Difference method show that the naïve group also had larger *Syl2_MaxDiff* than the trained vocalists (diff = 35.567, $p < .05$) and the L2 learners (diff = 25.928, $p < .05$).

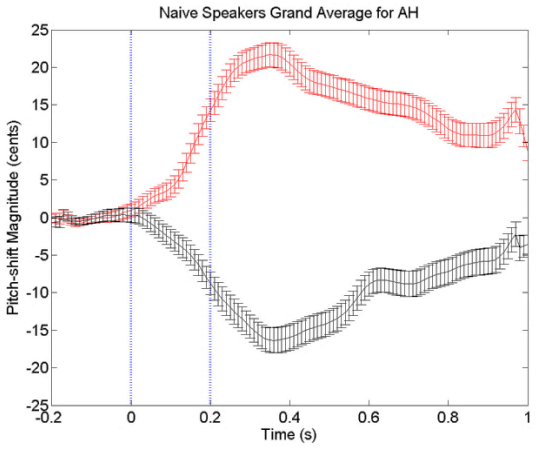
For the T14 comparison, no significant GROUP effect was found on the *Syl1_MeanDiff* ($F(3,32) = 0.815$, $p = .453$), *Syl2_MeanDiff* ($F(3,32) = 1.390$, $p = .258$) and the *Syl2_MaxDiff* ($F(3,32) = 1.004$, $p = .459$), which means the four language groups did not show differences between the compensatory responses to up-shift stimuli and down-shift stimuli on both /ma/ syllables.

4.3.5.2 Sustained vowel

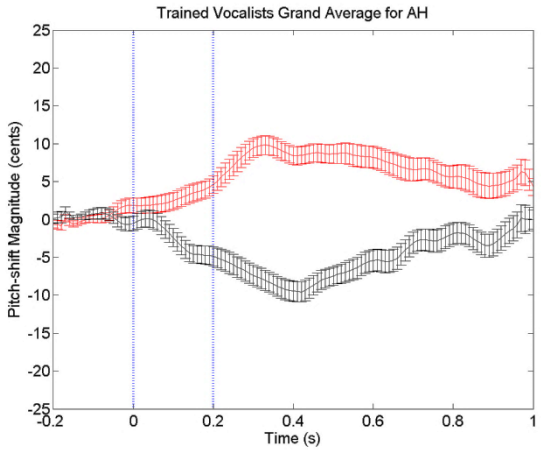
In the sustained vowel condition, I conducted separate permutation tests with repeated measures ANOVA fitting the factors of GROUP and DIRECTION (up-shift or down-shift) for each dependent variable (*onset time*, *peak time*, and *peak amplitude*), to investigate whether pitch-shift responses are affected by language experience or vocal training experience. The pitch-shift responses are displayed in Figure 4.8.

Figure 4.8 Pitch-shift responses during /a/ production by group.

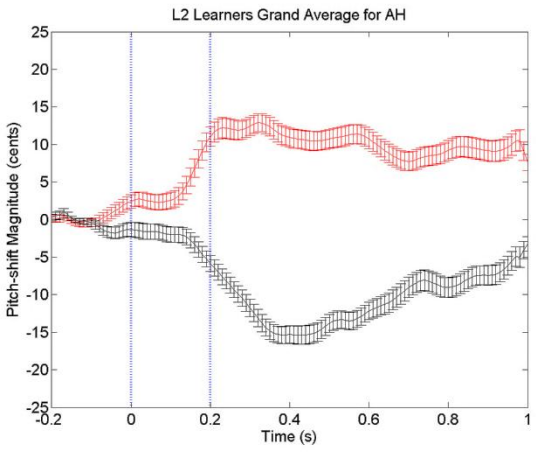
a.



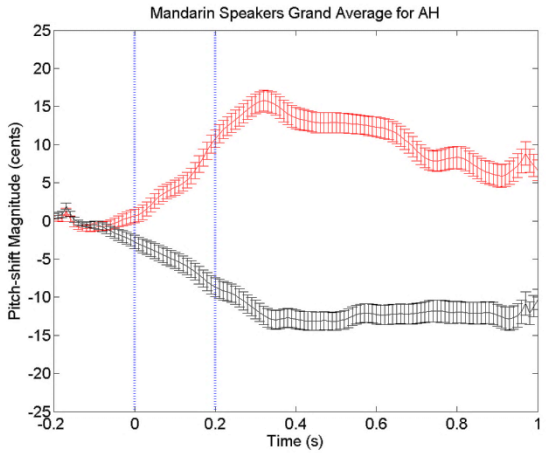
b.



c.



d.



Note. The x-axis represents time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli.

Black curves represent the responses to up-shift stimuli. The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

For *onset time*, there was no significant main effect of GROUP ($F(3,64) = 1.906$, $p = .114$) and of DIRECTION ($F(1,64) = 0.203$, $p = .660$). The interaction between GROUP and DIRECTION was not significant ($F(3,64) = 0.436$, $p = .881$). For *peak time*, no significant main effect of GROUP ($F(3,64) = 0.943$, $p = .632$) and of DIRECTION ($F(1,64) = 2.049$, $p = .327$) were found. The interaction between GROUP and DIRECTION was not significant ($F(3,64) = 0.260$, $p = .833$). For *peak amplitude*, absolute values of response peak amplitude were used in the ANOVA in order to examine the DIRECTION effect. There was a significant main effect of GROUP ($F(3,64) = 8.391$, $p < .0001$) but the main effect of DIRECTION ($F(1,64) = 2.037$, $p = .291$) and the interaction between GROUP and DIRECTION ($F(3,64) = 1.054$, $p = .388$) were not significant. Post-hoc comparisons with the Tukey's Honest Significant Difference method show that the naïve speakers had larger response amplitude for /a/ than the trained vocalists (diff = 10.308, $p < .001$) and Mandarin speakers (diff = 6.143, $p < .05$). The L2 learners had larger response amplitude for /a/ than the trained vocalists (diff = 7.331, $p < .01$).

4.3.5.3 Correlation between /a/ and Mandarin bi-tonal sequences

I compared the absolute peak amplitude of /a/ (up-shift and down-shift) and the *Syl2_MeanDiff* and *Syl2_MaxDiff* of T11, T12, and T14. The distribution-free spearman correlation analyses show that the absolute peak amplitude of /a/ in the up-shift condition was significantly correlated with the *Syl2_MeanDiff* of T11 ($r = 0.396$, $p < .01$). The absolute peak amplitude of /a/, either up-shift ($r = 0.540$, $p < .001$) or down-shift ($r = 0.280$, $p < .05$), was also significantly correlated with the *Syl2_MaxDiff* of T11. No significant correlations were found between absolute

peak amplitudes of /a/ and the amplitude deviations (*Syl2_MeanDiff* and *Syl2_MaxDiff*) of T12 and T14. This suggests that a smaller response peak amplitude in /a/ is correlated with smaller deviation between up-shift and down-shift responses in T11.

4.3.5.4 Correlation between perception and production

To further investigate whether performances in the nonlinguistic and Mandarin tone discrimination tasks were correlated with the pitch-shift responses, distribution-free spearman correlation analyses were conducted. The variables of the perception tasks included TD1 score (the first attempt), TD2 score (the second attempt), and MD overall accuracy. The variables of pitch-shift responses included T11*_Syl2_MeanDiff*, T11*_Syl2_MaxDiff*, T12*_Syl2_MeanDiff*, T12*_Syl2_MaxDiff*, T14*_Syl2_MeanDiff*, T14*_Syl2_MaxDiff*, *Vowel_onset time*, *Vowel_peak time* and *Vowel_peak amplitude*. The results show that MD accuracy was negatively correlated with T11*_Syl2_MeanDiff* (r = -0.443, p < .01), T11*_Syl2_MaxDiff* (r = -0.517, p < .01), T12*_Syl2_MeanDiff* (r = -0.362, p < .05), and T12*_Syl2_MaxDiff* (r = -0.387, p < .05), meaning that better MD accuracy was correlated with smaller deviations between up-shift and down-shift responses in T11 and T12. No other correlations were significant.

4.4 Discussion

In this study, I investigated whether tonal language experience or vocal training experience influences auditory discrimination and sensorimotor integration in the contexts of tonal language syllables or simple vowels. As a specific study of Mandarin language, this investigation also assesses whether specific attributes of tones, such as the high sustained tone, is differentially influenced by experience or learning. Regarding the findings for the nonlinguistic tone frequency discrimination task, I did not find statistical differences between naïve speakers, trained vocalists,

L2 learners, and Mandarin speakers. When Mandarin tone discrimination was compared, Mandarin speakers demonstrated significantly better performance compared to the other three groups. As for sensorimotor integration which was investigated with the pitch-shift responses, naïve speakers showed significantly larger deviations in vocal F0 compared to Mandarin speakers when they produced bisyllabic utterances with either the first or second tone in Mandarin (/ma1 ma1/ and /ma1 ma2/). It was also the naïve speakers who had the largest response peak amplitude during simple vowel production compared to trained vocalists and Mandarin speakers, while L2 learners had larger F0 response amplitudes for the simple syllable production relative to the trained vocalists. Two interesting associations were identified in this investigation between discrimination and pitch shift responses. First, participants with more accurate Mandarin tone discrimination task had smaller pitch-shift response ranges for tones. Secondly, these participants with smaller peak amplitude variation for tones also tended to have lower pitch shift response amplitudes for the vowel task. Altogether, these results point to an association between the ability to perceive tonal differences and attenuation of pitch-shift responses in the linguistic domain.

4.4.1 Attenuation of Pitch-shift Responses in Mandarin Speakers

This study is the first to examine linguistic specificity of pitch-shift responses within Mandarin speakers and non-native speakers with varying exposure to Mandarin. Xu et al. (2004) and Liu et al. (2009) investigated the susceptibility to pitch perturbation in Mandarin speakers and found inconsistent results. The Mandarin speakers' responses were never directly compared across Mandarin and non-Mandarin tasks or with speakers of other languages. The current study bridges this gap by directly comparing linguistic specificity within individual speakers and across language groups.

The results highlight that attenuation of pitch shift response amplitude is the dominant pattern among Mandarin speakers for both nonlinguistic (sustained vowel) and linguistic (Mandarin tone) domains, especially compared to naïve speakers. Therefore, this controlled study takes the initial report by Liu et al (2009) and places it in a broader context of pitch shift responses by Mandarin speakers. Although a mechanism for attenuation can only be speculated upon herein, the more neutral term attenuation resembles a suppression effect in which neural mechanisms actively suppress responses (which cannot be determined herein). The attenuation effect bears a certain similarity to Study 1 in which Mandarin speakers were less affected by magnitude and direction of the pitch-shift stimuli (/a/) than naïve speakers and L2 learners. Lower amplitude responses of Mandarin speakers suggest that their internal models of tone are different from naïve speakers'. Mandarin speakers may have more robust pitch control over self-produced vocalization and thus be less affected by auditory feedback deviations. I also remain cautious about over-interpreting the attenuation effect, because the pitch-shift amplitudes in this study were computed by taking the mean or maximum difference between responses to up-shift stimuli and responses to down-shift stimuli. The results might be different if other shift amplitudes had been studied (for instance, taking the difference between responses to up-/down-shift stimuli and responses to controls). However, I believe the measurements (mean or maximum difference) used in the current study are more informative than taking the difference between responses to up-/down-shift stimuli and responses to controls, because the controls recorded before the formal pitch-shifting section had more variance probably due to the fact that participants were trying to make themselves familiar with the task.

This study is also the first to relate linguistic specificity of pitch-shift responses to tone perception. The attenuation of pitch-shift responses in T11 and T12 was significantly correlated with the ability to perceive Mandarin tone contrasts, but not with the ability to perceive musical

tone contrasts. Participants with higher accuracy in Mandarin tone discrimination task tended to have smaller pitch amplitude deviations for T11 and T12 (*Syl2_MeanDiff* variable and *Syl2_MaxDiff* variable). However, a corresponding correlation was not observed between the ability to attenuate pitch-shift responses for the sustained vowel and the ability to discriminate Mandarin or musical tones. As the Mandarin speakers had highest linguistic tone discrimination accuracy, it suggests Mandarin speakers not only have more robust pitch control over self-produced vocalization but also have fine-tuned perception on linguistic tones. Lexical tone perception or internal model of tone is shaped by exposure to a tonal language through childhood language learning beginning in early childhood.

4.4.2 Second Language Tone Learning

The L2 learners in general did not show superiority in tone perception or pitch-shift attenuation that was comparable to the naïve speakers. Although L2 learners had attenuated pitch-shift responses in T12 compared to naïve speakers, no significant difference between the two groups was found for T11 or /a/. In terms of Mandarin tone discrimination, L2 learners did not show an advantage that resembled Mandarin speakers, because Mandarin speakers consistently had higher accuracy and further advancement than L2 learners. L2 learners did not even show a clear advantage relative to the other groups. Two of the L2 learners (participants 22 and 24; see Figure 4.3) even had degradation rather than advancement throughout the task.

This lack of advantage for L2 experience contrasts with the perceptual results of Study 1 in which L2 showed better perception than naïve speakers. Possible reasons for this different finding include the new task introduced herein that may have been more difficult. Additionally, this study did not solely involve L2 speakers currently enrolled in Mandarin classes as some of the L2 learners had taken Mandarin studies in the past. The lack of continuous learning could

have resulted in diminished performance on the Mandarin tone perception test. I still do expect that L2 learners are in a process of acquiring native-like internal models of tone, but either more consistent practice or further advancement in language learning might be necessary to detect evidence for changes within internal models. With more advanced students, native-like internal models of tone might be manifested both in pitch perception and voice F0 control. Testing students who either immersed in a Mandarin speaking environment or engaged in advanced studies is still predicted to indicate that pitch-shift responses are altered in a linguistic specific manner by language learning.

4.4.3. Vocal Training Experience

The current study included the innovative control condition of testing speakers with vocal training to determine if advanced levels of voice control might affect perception/production in a similar manner as language experience. Certainly, vocal training alters audio-vocal interactions (Burnett & Larson, 2002; Jones & Keough, 2008; Keough & Jones, 2009; Zarate, Wood, & Zatorre, 2010; Zarate & Zatorre, 2005, 2008). I found limited but relevant evidence that vocal training was associated with attenuated pitch shift amplitude for simple vowel production and the T12 condition. In both conditions, trained vocalists had pitch-shift response amplitudes that were attenuated compared to naïve speakers and L2 learners. This attenuation result for the sustained vowel corresponds to the important reports by Jones and Keough (2008) and Zarate and Zatorre (2005, 2008) where singers displayed attenuated pitch-shift responses. Following both of these reports, I also argue that singers rely more on internal models than non-singers in order to regulate voice F0. Unlike Mandarin speakers however, the trained vocalists did not show significant attenuation in T11 compared to naïve speakers. This difference suggests trained vocalists can attenuate the perturbation in certain contexts, but the ability does not fully extend to

all contexts. These singers did not have experience with Mandarin, so attenuation in the linguistic domain (T11) appears to require linguistically specific training. Therefore, I argue that two pathways to refinement or specialization of internal model are evident in the results. The first pathway is proficiency in a tonal language and the second is extensive vocal training. These two pathways do not result in common attenuation across all contexts however, so internal model refinement is again shown to be specific to training as argued by Keough and Jones (2009), where they found singers but not nonsingers were continually updating their internal models to adapt to gradual manipulations of altered feedback. In the current study, the difference in attenuation which distinguished these two groups was for T11. It is interesting that maintaining a high level tone in Mandarin was not identical to maintaining a sustained vowel and thus could require motor experience specific to language.

In terms of Mandarin tone perception, it was surprising that these highly trained vocalists did not perform better than naïve speakers. Two of the trained vocalists (participants 14 and 15; see Figure 4.3) did not make any advancement throughout the task. One possibility is that the trained vocalists were overly sensitive to the pitch differences which do not lead to tonal contrasts in Mandarin. Therefore, being able to discriminate lexical tones requires some linguistic tonal experience. However, one trained vocalist (participant 13) could perceive abstract tone (get to stair 7: different segments and different speakers' voices). Neither naïve speakers nor L2 learners were able to reach the highest level in the adaptive task. Thus, having some vocal training or musical training may still benefit tone learning to a certain extent. With Mandarin training, it is possible that trained singers might advance more rapidly in tone acquisition and accordingly show refinement of internal models earlier than L2 learners without voice training.

4.4.4. Is There Anything Special about the High Level Tone (T11) in Mandarin?

A general finding of this study is that pitch-shift response magnitudes for the sustained vowel /a/ were correlated with pitch-shift magnitude for T11, but not with T12 and T14. Participants who had smaller response peak amplitude to /a/ tended to have smaller *Syl2_MeanDiff* and *Syl2_MaxDiff* in T11. The sustained vowel and T11 tasks were similar in that a constant F0 was a goal. The other two linguistic conditions required rapid changes in F0 that could cover almost half an octave. It is very likely that detection of pitch shift responses is facilitated by tasks with less F0 variation. Additionally, regulation of a consistent F0 could be more subject to ongoing feedback than tasks which require a large and rapid change in F0.

Nevertheless, results in the T11 condition were not identical to the sustained vowel condition. For the sustained vowel /a/, both trained vocalists and Mandarin speakers attenuated the pitch perturbation and showed descriptively less F0 variation than other speaker groups. However, maintaining F0 for the high level tone (T11) in the linguistic domain was only associated with attenuation for Mandarin speakers, not for trained vocalists. The group specific discrepancy supports my argument that pitch-shift attenuation is influenced by specific language experience but most evidently for the high level tone.

Other language specific effects were observed in the T11 condition. Naïve speakers who had the largest pitch shift responses displayed clear compensation in response to the direction of the perturbation (see Figure 4.5a). Mandarin speakers' actual tone production appeared to differ from each of the other speaker groups as only Mandarin speakers had a declination in F0 from the first to second syllable (see Figure 4.5d). Shih (1997, 2000) has previously observed this prosodic characteristic in the speech of Mandarin. None of the other groups including the L2 learners introduced this feature in their production. Given that there were no instructions that indicated a declination was required or expected, this feature seems to have been a linguistic

feature that only Mandarin speakers would have considered appropriate. It is quite possible that any speech planning that involved attention to a prosodic feature in Mandarin speakers rendered their production of the T11 sequence less susceptible to the perturbation.

The marked group differences for the high level tone further stand out due to the relatively modest findings for the other two tones. For T12 (Figure 4.6), it was only naïve speakers who showed bidirectional compensatory responses. There was some indication Mandarin speakers and L2 learners attenuated pitch shift responses in a similar manner by reducing the depth of the F0 concave compared to the control trials during T12 production. The trained vocalists, in contrast, increased the depth of the concaves relative to control trials. So although some attenuation was evident in T12, it was less consistent than T11 across the groups.

Regarding T14, responses to either up-shift stimuli or down-shift stimuli overlapped almost entirely for each group (Figure 4.7). Statistically, no significant GROUP differences in the *Syl2_MeanDiff* and *Syl2_MaxDiff* were detected, suggesting that any perturbation response during the falling tone is more difficult to detect and more subtle than the high level tone. Perhaps, the compensation is more easily blocked out when speakers (irrespective of their background) abruptly decrease F0.

It appears that regulating the F0 of high level tone is challenging and requires some linguistic-specific experience. In phonological theories, whether contour tones are allophonic variants of level tones has been a matter of debate (Duanmu, 1994; Goldsmith, 1994; Newman, 1986). Because pitch-shift responses to the high level tone are different from pitch-shift responses to contour tones, I introduce the possibility that the underlying mechanism(s) for producing high level tone is distinct from that of contour tones. In other words, high level tone production in Mandarin appears to be managed in a unique manner by native speakers that is not apparent in L2 learners or other naïve speakers who are imitating this tone.

Fujisaki's command-response model for generating F0 contours may give some insight into the different effects of pitch shift responses on high level tones versus contour tones (Fujisaki & Hirose, 1984; Fujisaki, Ohno, & Gu, 2004). The mathematical model was originally built to account for the accent and intonation in Japanese, where the phrase commands are a set of impulses and the accent commands are a set of stepwise functions. Intonation is then generated by superimposing the phrase commands and accent commands and then applying a logarithmic function (Fujisaki & Hirose, 1984). The command-response model was extended to tonal languages, including Thai, Cantonese, and Mandarin Chinese (Chomphan & Chompunth, 2012; Fujisaki, Wang, Ohno, & Gu, 2005; Gu, Hirose, & Fujisaki, 2007). In Mandarin, the phrase commands are composed of impulses, controlling the global shape of F0 contours (intonation), while the tone commands are pedestal functions varying in polarity with lexical tone contours. In Fujisaki's mathematical model, tone 1 has positive tone commands lasting about 0.5 second, tone 2 has negative plus positive tone commands lasting about 0.5 second, and tone 4 has positive plus negative tone commands lasting about 0.25 second (Fujisaki et al., 2005). As tone 1 is inherently long, the positive tone commands should be continuously executed to maintain the high pitch. The rising part of tone 2 has the same requirement to keep increasing the pitch. In tone 4, the positive command is placed early, so the negative command is already in effect when pitch perturbation occurs. Tone 4 is the only case where pitch perturbation happens in a negative command. The requirement of continuously executing positive tone commands in tone 1 (or even tone 2) may lead to the fact that attenuating pitch perturbation in high (level) tone is harder than the falling tone (tone 4).

4.4.5. Nonlinguistic Tone Discrimination

My expectation that nonlinguistic tone discrimination might be influenced by the listeners'

language or vocal training experience was not supported. This comparable group performance stands in sharp contrast to Study 1, in which L2 learners and Mandarin speakers showed more accurate tone discrimination than naïve speakers. An important procedural difference between the two studies is that the accuracy feedback shown on the screen was covered (correct or wrong for each trial) in the current study when participants were doing the nonlinguistic tone discrimination task. The purpose of concealing accuracy feedback was to avoid a learning effect. It is possible that L2 learners and Mandarin speakers in Study 1 could detect pure tone differences more effectively when prompt feedback was delivered following the response. However, when the feedback is removed, neither L2 learners nor Mandarin speakers displayed an advantage.

Research has shown that musicians have better pitch discrimination accuracy than nonmusicians (Koelsch, Schroger, & Tervaniemi, 1999; Tervaniemi, Just, Koelsch, Widmann, & Schroger, 2005), but I did not find an advantage in the trained vocalists. However, from the post-hoc interviews, I found that participants with a longer period of musical training had better performance on nonlinguistic tone discrimination. The duration of musical training was only related to nonlinguistic tone discrimination, not to Mandarin tone discrimination, suggesting pitch perception is domain-specific.

4.4.6. Significance of the Study

Controlling voice F0 in syllable domain is a language-specific skill that Mandarin speakers acquire in early childhood. The current study shows that this particular motor skill is associated with the ability to perceive tonal differences. In the terminology of internal model theory, Mandarin speakers' language internal models enable stable tone production even when potentially perturbing stimuli are presented. This is one hallmark of mature perceptual and motor

systems that is not apparent in second language learners. Although trained vocalists have the potential to produce difficult tonal contours, regulation of voice F0 in a linguistic domain may still require specific tone training. It remains to be shown whether trained vocalists are able to transfer the motor skill from nonlinguistic domain to the linguistic domain faster than L2 learners without vocal musical training.

4.5 Conclusion

I examined both perception and audio-vocal responses in four different groups, including naïve speakers, trained vocalists, L2 learners and Mandarin speakers. Listeners with tonal language experience or vocal training experience were not superior in nonlinguistic tone discrimination to those without. However, the ability to perceive linguistic tone contrasts is affected by language experience, with Mandarin speakers outperforming the other groups. In terms of audio-vocal responses, speakers with tonal language experience or vocal training experience could attenuate pitch-shift responses in a similar fashion, and the attenuation effect was stimulus-specific:

i) for /a/, Mandarin speakers and trained vocalists had smaller response peak amplitudes than naïve speakers,

ii) for the high level tone (T11), only Mandarin speakers attenuated the pitch-shift responses and maintained the declination effect,

iii) for the rising tone (T12), Mandarin speakers and L2 learners attenuated the pitch-shift responses by reducing the depth of the tonal concaves

iv) for the falling tone (T14), all four groups suppressed the pitch-shift responses.

This suggests that the high level tone is special because suppressing the responses to the high

level tone requires mastery of tonal languages. The results support the hypothesis that Mandarin speakers have more robust internal models for tone, as their ability to control F0 is in general (in both linguistic and nonlinguistic domains) superior to naïve speakers, trained vocalists and L2 learners. The L2 learners showed limited resemblance to Mandarin speakers in terms of perception and pitch-shift response attenuation suggesting more tone training in required to achieve native-like internal models.

CHAPTER 5
STUDY 3: THE EFFECT OF REAL-TIME VISUAL FEEDBACK ON SENSORIMOTOR RESPONSES

5.1 Introduction

Study 1 and Study 2 investigated two external factors related to pitch processing, including language experience and musical experience. The results that show suppressed responses to pitch perturbation indicate that Mandarin speakers appear to have more stable internal models of F0 control than non-tone speakers. Trained vocalists also had suppressed responses to pitch perturbation for sustained vowels but not for all Mandarin tones. Study 3 will continue the investigation of pitch-shift response suppression by investigating another external factor that may influence F0 control. Real-time visual feedback may influence pitch-shift responses because a visual signal showing continuous F0 can be used to stabilize F0, which has been confirmed in singing (Ferguson, Moere, & Cabrera, 2005; Howard et al., 2007; Howard et al., 2004; Thorpe, 2002; Wilson et al., 2008).

In motor learning, it has been argued that an effective way to enhance motor skills is to direct learners to use an external focus, such as the motion of golf club in golf learning, or the markers on the stabilometer platform in balance learning, rather than using an internal focus, such as the swing of the arm in golf learning or paying attention to the feet in balance learning (Lauterbach, Toole, & Wulf, 1999; Peh, Chow, & Davids, 2011; Shea & Wulf, 1999; Wulf, Höß, & Prinz, 1998; Wulf, Shea, & Lewthwaite, 2010). Visual feedback is a type of external focus that may influence motor control. It is possible that in tone/pitch production, re-directing speakers' attention to pitch tracking results rather than instructing the speaker to ignore pitch perturbations could facilitate voice F0 stability (i.e., with attenuated pitch-shift responses). Study 3 will test this question by providing real-time visual feedback of continuous F0 to assess whether speakers

show greater suppression of pitch-shift responses compared to auditory only conditions.

The contribution of real-time visual feedback of F0 (or other voice features) has been explored in singing instruction. Singers who received visual feedback of acoustic voice features showed better pitch accuracy than those taught by a traditional method (verbal instruction) (Ferguson et al., 2005; Howard et al., 2007; Howard et al., 2004; Thorpe, 2002; Wilson et al., 2008). Nonmusicians who received visual feedback related to musical note frequencies had better pitch recognition than control participants who did not receive training with visual feedback (Eldridge, Saltzman, & Lahav, 2010).

In terms of language learning, L2 learners of Chinese showed improvement in tone production by visualizing pitch contours. L2 learners of French were able to generalize intonation to novel sentences with the help of visible pitch contours (Chun, Jiang, & Avila, 2013; Hardison, 2004; Levis & Pickering, 2004). These results are consistent with a cross-modal advantage in vocal learning. The term cross-modal refers to a contribution of the visual modality to tasks, which are typically dependent on the auditory signal for learning.

Findings from Study 2 show that pitch-shift responses during sustained vowels differ from pitch-shift responses during lexical tones. The sensitivity of pitch-shift responses supports the view that pitch-shift responses are stimulus-specific. This specificity corresponds to previous research's results in which pitch-shift response magnitudes were larger in singing than in speaking (Natke et al., 2003) and in speech (question intonation) compared to single vowels (S. H. Chen et al., 2007).

The general trend is complicated because Zarate et al. (2010) showed that singers were less able to suppress pitch-shift responses to 25 cents stimuli than to 200 cents stimuli, suggesting that responses to smaller shifts are under less voluntary control. While this finding supports claims that pitch-shift responses are tuned to the correction of small pitch perturbation (Hain et

al., 2000; Liu & Larson, 2007), it means that attempts to understand pitch-shift response suppression must account for multiple factors.

In Study 3, different routes to pitch-shift suppression will be contrasted by varying perturbation magnitude, linguistic specificity and visual feedback. This study is necessary to determine whether pitch-shift responses suppression is a general mechanism or specific to particular manipulations. I hypothesize that real-time visual feedback of voice F0 could enable both Mandarin speakers and non-tone speakers (naïve speakers) to stabilize their voice F0 in the presence of pitch perturbations, but the degree of suppression will vary as the linguistic context and perturbation magnitude are manipulated. One possible outcome is that visual feedback will lead to similar suppression despite variation in linguistic context and perturbation condition. If so, pitch-shift response suppression may be too general and of limited utility for the study of language. On the other hand, if there are differences in the degree of suppression depending on the stimulus, perturbation magnitude or feedback, then the pitch-shift paradigm may be a specialized tool for probing F0 control and learning. One possibility is that non-tone speakers will show similar suppression of pitch-shift responses across the linguistic manipulation, while Mandarin speakers may show an additive suppression with visual feedback leading to even greater suppression when producing lexical tones. Another possibility is that visual feedback will cause pitch-shift response suppression for small (25 cent) perturbations, which would indicate that even the more involuntary aspects of the pitch shift response are liable to modulation and highly useful for probing audio-vocal control.

5.2 Experiments in Study 3

The two tasks conducted in Study 3 are presented below. The pitch-shift task investigates whether sensorimotor control over voice F0 is dependent on language experience and is subject

to feedback mode. The nonlinguistic tone discrimination task examines the participants' ability to discriminate two tones with different fundamental frequency. The same participants participated in both tasks.

5.2.1 Participants

Twenty participants participated in the experiment including 10 naïve speakers (1 male) who were never exposed to tonal languages and 10 native speakers of Mandarin (5 male). Both groups included some participants with musical training before college, but none possessed ongoing participation in formal musical training. Based on the post-hoc interviews, the average period of musical training (instruments) before college was 4 ± 1.2 years for naïve speakers and 6.6 ± 2.0 years for Mandarin speakers. All were recruited from University of Illinois at Urbana-Champaign (age: 20-35 years old). All participants passed a hearing screening at 20 dB HL bilaterally at 125, 250, 500, 750, 1000, 2000, 3000, and 4000 Hz. None of the participants reported a history of neurological or communication disorders. They were paid $10 for their participation. All methods reported herein were approved by the Institutional Review Board at the University of Illinois and all participants provided written informed consent.

5.2.2 Pitch-shift Task

5.2.2.1 Materials

There were 2 (LING: /a/ or /ma1/) x 2 (MODE: AUDIO-ONLY or AUDIO-VISUAL) x 2 (MAGNITUDE: 25 cents or 200 cents) experimental conditions in total. The order of the 8 experimental sections is presented in (3).

(3) Pitch-shift stimuli and conditions in Study 3

a. AUDIO-ONLY /a/ 200 cents

b. AUDIO-ONLY /a/ 25 cents

c. AUDIO-ONLY /ma1/ 200 cents

d. AUDIO-ONLY /ma1/ 25 cents

e. AUDIO-VISUAL /a/ 200 cents

f. AUDIO-VISUAL /a/ 25 cents

g. AUDIO-VISUAL /ma1/ 200 cents

h. AUDIO-VISUAL /ma1/ 25 cents

In the AUDIO-ONLY MODE, participants were given auditory feedback only. In the AUDIO-VISUAL MODE, participants were given real-time visual feedback of their vocal F0 in addition to their own auditory feedback.

As for the MAGNITUDE of pitch-shift stimulus, at a randomized interval (1000, 1100, and 1200 ms) after vocalization onset, the pitch of the feedback signal was altered for 200 ms by either a ±25 cent shift or a ±200 cent shift with shift DIRECTION randomized. Forty-five vocalizations were recorded with one pitch-shift perturbation per trial in each condition, with the number of upward shifts, downward shifts and control stimuli (no-shift) approximately equal.

5.2.2.2 Procedures

Participants were seated in a sound booth and wore Sennheiser HD-280 Pro headphones and a headworn Shure microphone placed 2 cm away from the corner of the mouth. They were trained to produce the vowel /a/ and the Mandarin word /ma1/ (“mother” with a high level tone) at approximately 70 dB sound pressure level (SPL) while self-monitoring their vocal volume on

a Dorrough Loudness Monitor (model 40-A) placed in front of them. The voice signal from the microphone was amplified with a YAMAHA mixer (MG102c) and sent to an Eventide Ultra-Harmonizer (model H7600) that generated pitch shifts. The participant's own voice signal was amplified with a Samson S-phone Headphone Amplifier and played back to him/her at approximately 80 dB SPL to reduce the possible influence of bone conduction. MIDI software (MAX/MSP v.5 by Cycling 74, Walnut, CA) connected to a MOTU (model UltraLite-mk3 Hybrid) controlled the timing, duration and magnitude of the pitch shifts via the Eventide ultra-harmonizer. The participant's voice, altered feedback signal, and pitch-shift events (transistor transistor logic (TTL) pulses) were digitized at 5 kHz per channel with WINDAQ/Pro software (DATAQ Instruments, Inc., Akron, OH). The stimuli (the vowel /a/ and the character of the Mandarin word /ma1/ ("mother")) were shown on a Dell laptop with E-prime software (v.1, Psychology Software Tools, Inc., Sharpsburg, PA).

In the sustained vowel condition, participants first heard a male /a/ sound recording which lasted for 3 seconds. They had to repeat the /a/ vocalization for 3 seconds after the vowel stimulus ended. In the Mandarin condition, participants first heard a female /ma1/ sound recording which was 2 second long. Then they repeated or imitated the Mandarin sound and held it for 3 seconds.

In the AUDIO-ONLY MODE, participants were instructed to hold vocal pitch and volume as steady as possible and to ignore any changes in what they heard. In the AUDIO-VISUAL MODE, the real-time visual feedback on pitch was tracked and displayed by Computerized Speech Lab (CSL™) (Model 4500, KayPENTAX, Montvale, NJ). Participants were given two reference lines of F0 (their mean F0 $\pm$ 5 Hz) on the pitch display. They were instructed to correct their voice immediately if the pitch-tracking results showed their voice was not steady enough or drifted above or below the reference lines.

Speakers may learn to stabilize their voice with the help of visible F0 contours and apply the learned skill when the visual feedback is not provided. In order to avoid the learning effect, AUDIO-ONLY conditions preceded AUDIO-VISUAL conditions. A short practice section (5 trials) was given before the experimental sections (3a), (3c), (3e), and (3g) started.

5.2.3 Nonlinguistic Tone Discrimination Task

In addition to the pitch-shift task, participants' nonlinguistic tone discrimination was examined to see if the perception is correlated with the ability to suppress pitch-shift responses. Following the method described for Study 2, the Adaptive Pitch Test (Mandell, 2009) on the tonometric website was presented via laptop to each participant. The intensity was adjusted to each subject's most comfortable level. Participants listened to a series of two short tones and were asked whether the second tone was lower or higher in pitch than the first tone. Accuracy feedback (correct or wrong for each trial) was covered by paper attached to the screen. The task took about 3 minutes to complete.

5.2.4 Data Analysis

In the pitch-shift task, the raw signals recorded in WINDAQ were imported into MATLAB (R2009b, The Mathworks Inc., Natick, MA). The signals were sorted into files based on the direction of pitch shift stimuli (up-shift, down-shift, or no-shift). For each vocalization, a 1.2 second window including a 200 ms pre-pulse baseline, the 200 ms pulse, and an 800 ms post-pulse period were selected for acoustic analysis. The *pitch values* were tracked using the SWIPE pitch estimation algorithm at 10 ms intervals (Camacho & Harris, 2008) and then converted into cents (cents = $1200*\log_2(F0/\text{baseline } F0)$), where the baseline F0 was the mean F0 between 0-200 ms. Following responses and responses without a steady baseline (the phase

before pitch-shift occurs) were excluded in this study so that only compensatory responses were considered.

The onset time, peak time, and peak amplitude of the average responses (see Figure 3.2 on p.56) were obtained in the following way: *onset time* was defined as the time at which the response exceeded 2 standard deviations of the pre-stimulus mean and the time was at least 60 ms after the pitch-shift stimulus onset; *peak time* and *peak amplitude*, which corresponded to the first peak of averaged response, usually occurred between 200-600 ms after the onset of the pitch-shift stimulus and were selected with a peak-picking algorithm in MATLAB. In addition to *peak amplitude*, the relative gain of the response (*relative gain*) was computed by dividing the response magnitude by the stimulus magnitude (either 25 cents or 200 cents) and multiplying by 100 %.

A one-way ANOVA was performed to investigate the GROUP differences in nonlinguistic tone discrimination. To evaluate pitch-shift responses, repeated measures ANOVAs were fitted to the linear model of GROUP, LING, MODE, DIRECTION and MAGNITUDE and conducted separately for each dependent variable *onset time*, *peak time*, *peak amplitude* and *relative gain*. Correlations were conducted to examine whether pitch-shift responses were related to perception performance.

As in Study 1, the responses of the whole F0 contours to pitch-shift perturbation were also modeled by using Generalized additive models GAMs (Wood, 2006). A series of model comparisons was conducted to investigate whether inclusion of a predictor (independent variable) leads to a significantly better fit of the model to the absolute F0 records (dependent variable). The F0 records fitted into the GAM were the *pitch values* obtained from the pitch tracking algorithm in Praat at 10 ms interval which were then converted into cents. First, SUBJECT and TIME treated as random effects, `s(SUBJ,bs="re")` and `s(TIME,bs="re")`, were included

in the model, which serves as a baseline. Then, the predictors were included one at a time in the following order: GROUP, temporal smoothing with restricted cubic splines for GROUP, LING, temporal smoothing with restricted cubic splines for LING, MODE, temporal smoothing with restricted cubic splines for MODE, DIRECTION, temporal smoothing with restricted cubic splines for DIRECTION, MAGNITUDE, and temporal smoothing with restricted cubic splines for MAGNITUDE. Significance of parametric terms (GROUP, LING, MODE, DIRECTION and MAGNITUDE) is evaluated by means of the t-tests, while significance of nonparametric terms (smooth terms) is evaluated by means of Bayesian *p*-values.

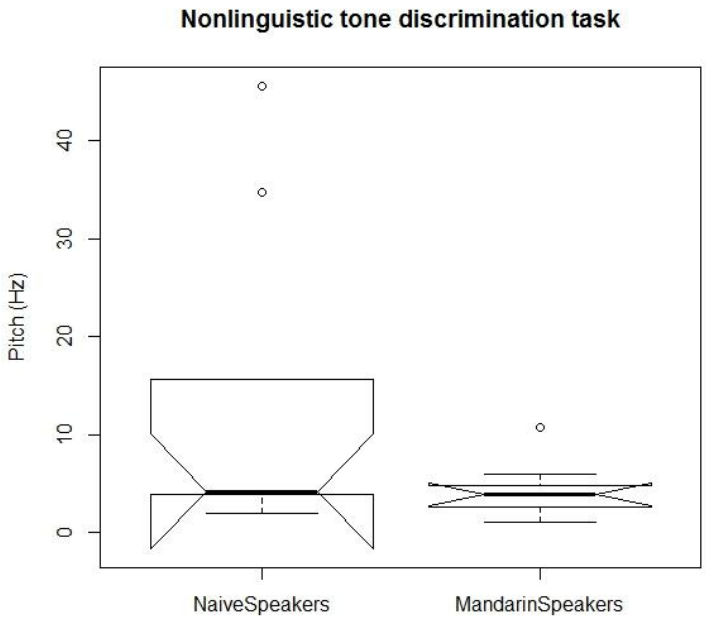
5.3 Results

This section presents the results of nonlinguistic tone discrimination task and pitch-shift task. Since the perception data and production data violated the assumptions of normal distribution and homogeneity of variance, permutation tests were used to examine statistical significance.

5.3.1 Nonlinguistic Tone Discrimination

The scores on the nonlinguistic tone discrimination task (TD) are depicted in Figure 5.1. In the nonlinguistic tone discrimination task, a lower score in Hz corresponds to better discrimination ability. The permutation test with ANOVA shows that there was no significant main effect of GROUP ($F(1,18) = 2.662$, $p = .111$) on the nonlinguistic discrimination scores.

Figure 5.1 Discrimination performance in the nonlinguistic tone discrimination task.



5.3.2 Pitch-shift Task

Only compensatory responses with steady baselines were included in the data analysis, excluding 33% of the trials overall which were composed of either following responses or non-responses. In any case, following responses constituted a minority of responses in each participant. The pitch-shift responses of naïve speakers and Mandarin speakers are displayed in Figure 5.2 and Figure 5.3, respectively. As response latencies and response amplitudes are not meaningful for the control stimuli (no shift), vocal responses to the controls were excluded in the statistical analyses.

Figure 5.2 Naïve speakers' pitch-shift responses.

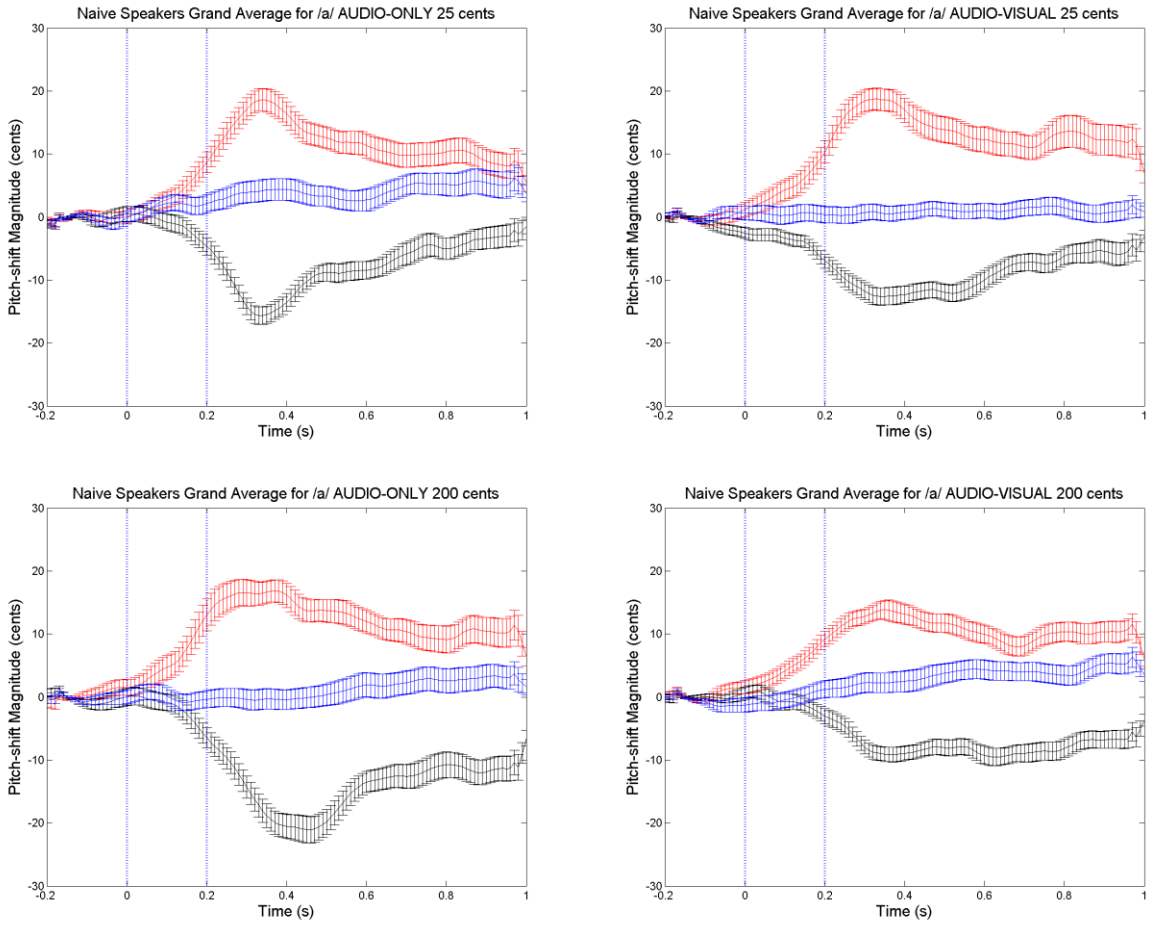
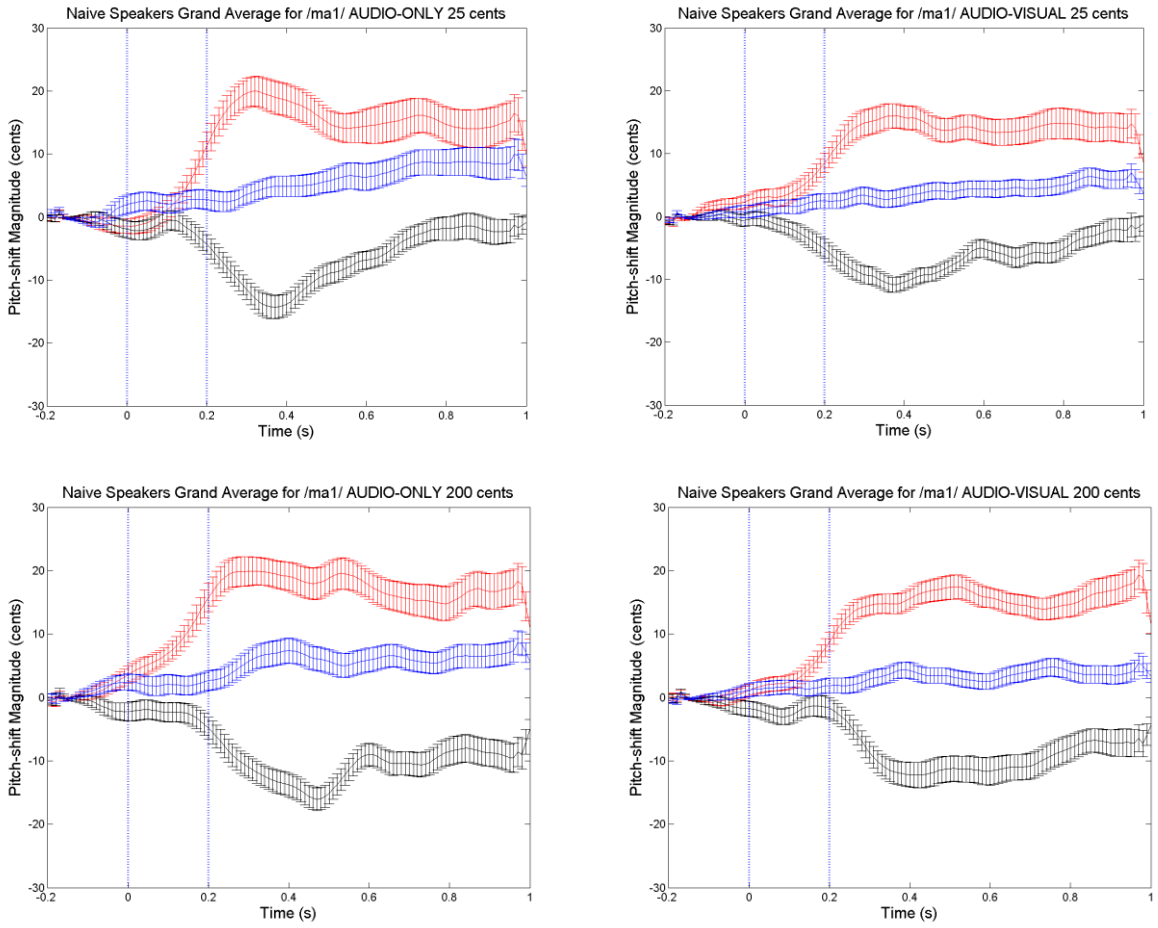


Figure 5.2 (Cont.)



Note. The x-axis is time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Black curves represent the responses to up-shift stimuli. Blue curves represent the responses to the controls (no shift). The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

Figure 5.3 Mandarin speakers' pitch-shift responses.

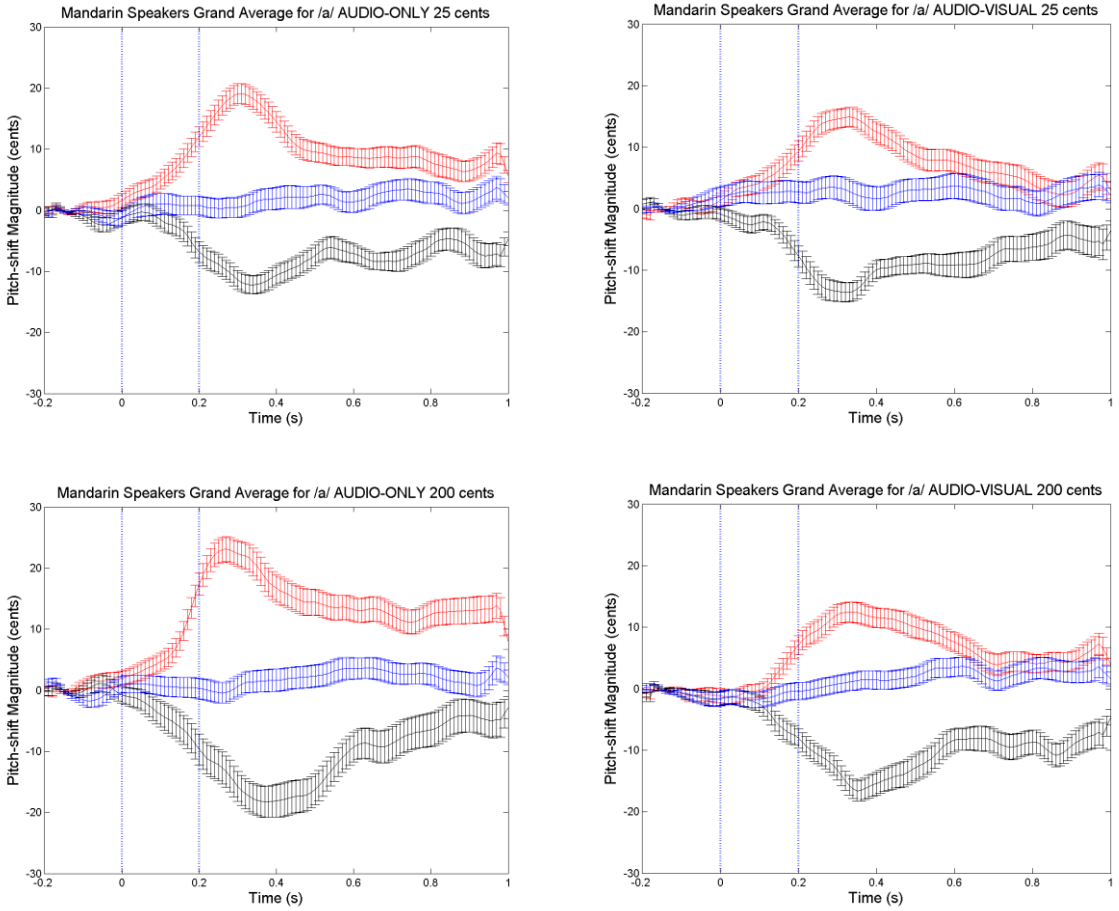
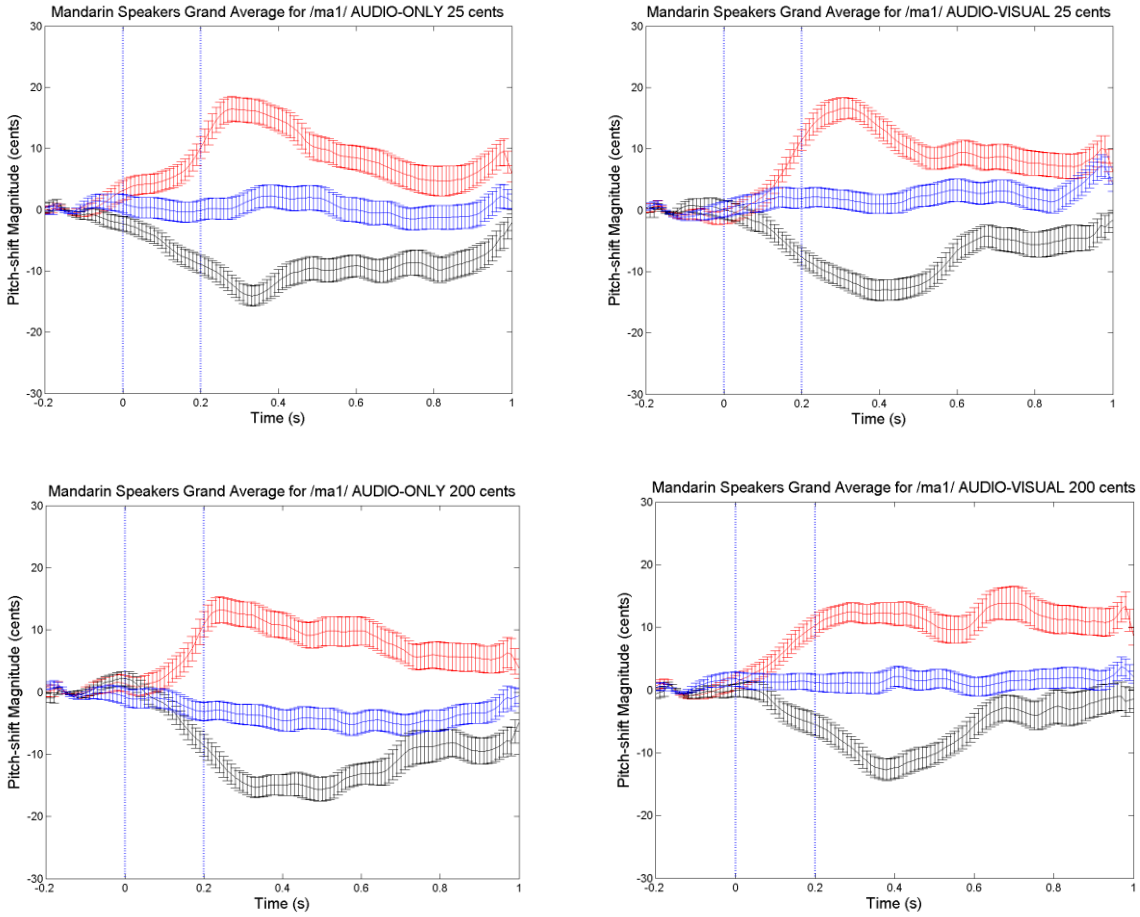


Figure 5.3 (Cont.)



Note. The x-axis is time measured in second. The y-axis represents the pitch-shift response magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Black curves represent the responses to up-shift stimuli. Blue curves represent the responses to the controls (no shift). The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

For *onset time*, there was a significant main effect of DIRECTION ($F(1,314) = 5.997$, $p < .0001$), with response onset time being faster in the down-shift conditions (150.625 ± 5.476 ms) than in the up-shift conditions (171.438 ± 6.502 ms) (**DOWN < UP**). No significant main effects of GROUP ($F(1,314) = .100$, $p = .961$), LING ($F(1,314) = .552$, $p = .961$), MODE ($F(1,314) = 1.045$, $p = .556$), and MAGNITUDE ($F(1,314) = 2.454$, $p = .135$) on onset time were found.

For the *peak time*, there were significant main effects of GROUP ($F(1,314) = 11.644$, $p < .0001$), DIRECTION ($F(1,314) = 34.815$, $p < .0001$), and MAGNITUDE ($F(1,314) = 3.978$, $p < .05$). Mandarin speakers (333.188 ± 6.574 ms) had shorter peak times than naïve speakers (363.875 ± 6.824 ms) (**MANDARIN < NAIVE**). Response peak time was also faster in the down-shift condition (322 ± 6.134 ms) than in the up-shift condition (375.062 ± 6.801 ms) (**DOWN < UP**), and also in the 25 cents condition (339.563 ± 5.868 ms) than in the 200 cents condition (357.500 ± 7.569 ms) (**25 CENTS < 200 CENTS**). No significant effects of LING ($F(1,314) = .217$, $p = .541$) or MODE ($F(1,314) = .004$, $p = 1.000$) for peak time were found.

For the *peak amplitude* and *relative gain* variables, I compared the absolute peak amplitudes of all conditions. There was a significant main effect of MODE ($F(1,314) = 15.417$, $p < .0001$) for peak amplitude. Response peak amplitude was smaller in the audio-visual condition (16.138 ± 0.566 cents) than in the audio-only condition (19.697 ± 0.715 cents) (**AUDIO-VISUAL < AUDIO-ONLY**). No significant main effects of GROUP ($F(1,314) = .106$, $p = 1.000$), LING ($F(1,314) = 1.939$, $p = .686$), DIRECTION ($F(1,314) = 5.392$, $p = .068$), and MAGNITUDE ($F(1,314) = .559$, $p = .229$) were detected for peak amplitude. For the relative gain, there were significant main effects of MODE ($F(1,314) = 4.390$, $p < .05$), DIRECTION ($F(1,314) = 4.375$, $p < .001$), and MAGNITUDE ($F(1,314) = 490.328$, $p < .0001$), with relative gain being smaller in the audio-visual condition (36.826 ± 2.859 %) than in the audio-only condition (42.615 ± 3.356 %) (**AUDIO-VISUAL < AUDIO-ONLY**), smaller in the up-shift condition (36.831 ± 2.892 %) than in the down-shift condition (42.610 ± 3.327 %) (**UP < DOWN**), and smaller in the 200 cents condition (9.128 ± 0.313 %) than in the 25 cents condition (70.313 ± 2.770 %) (**200 CENTS < 25 CENTS**). No significant main effects of GROUP ($F(1,314) = .516$, $p = .190$) and LING ($F(1,314) = .275$, $p = .667$) were found for relative gain.

5.3.3 Correlation between Perception and Production

To further investigate whether performance in the nonlinguistic task was correlated with the pitch-shift responses, distribution-free spearman correlation analyses were conducted; however, no significant correlations were found.

5.3.4 Generalized Additive Models: Modeling the Pitch Values in the Pitch-shift Task

Generalized additive models (GAM) were fitted with the absolute F0 values as the dependent variable. As in the ANOVA analyses, F0 contours for the control stimuli were excluded. GAM analysis shows that there were no significant differences in the F0 contours before perturbation onset (200 ms long), providing one indication that significant drift or variation in F0 was not affecting the F0 in the pre-perturbation periods. To examine how speakers respond to pitch-shift stimuli, only the F0 contours after the onset of perturbation (1 second long for each curve) were considered in GAMs. The sequence of model comparisons is summarized in Table 5.1. Each row compares two models, where the second model has one more predictor or interaction term than the first model. The evaluation was based on whether there was a reduction in deviation (Akaike Information Criterion, AIC for short), and whether this reduction was significant given the effective degrees of freedom (*edf*). Significance was evaluated with an F test. The baseline model, which is not shown in the table, included SUBJECT and TIME as random effects. The first row indicates that including the predictor GROUP reduced the AIC by -0.007. The *F*-test shows that inclusion of GROUP as a predictor lead to a significantly better fit of the model to the data ($p < 0.05$). However, significant differences between the groups developed as time progressed (reduction in AIC 5807.724, $p<.0001$). The model was improved when the predictors LING, MODE, DIRECTION and MAGNITUDE, and their interactions with TIME were included, which significantly reduced the AIC.

Table 5.1. Generalized Additive Model: Sequential model comparison in Study 3

Predictor	*edf*	Reduction AIC	*F*	*p*
GROUP	1	-0.007	1.000	<.05
s(TIME, GROUP)	17.170	5807.724	394.97	<.0001
LING	1	28.499	30.832	<.0001
s(TIME, LING)	6.800	92.486	18.328	<.0001
MODE	1	432.087	423.36	<.0001
s(TIME, MODE)	7.002	98.510	18.694	<.0001
DIRECTION	1	930.781	880.38	<.0001
s(TIME, DIRECTION)	15.101	270.764	36.941	<.0001
MAGNITUDE	1	52.784	54.276	<.0001
s(TIME, MAGNITUDE)	10.800	122.284	15.236	<.0001

The final model presented in Table 5.2 indicates the GROUP, LING, MODE, DIRECTION, and MAGNITUDE had significant influences on F0 changes over time. Since the interactions between TIME and GROUP, between TIME and LING, between TIME and MODE, between TIME and DIRECTION, and between TIME and MAGNITUDE were all significant, stratified post-hoc comparisons were performed on the coefficients of 2 GROUPs, 2 LINGs, 2 MODEs, 2 DIRECTIONs and 2 MAGNITUDEs. Significance was evaluated with a Wald Chi-squared test for model coefficients. For GROUP, there were significant differences in F0 contours between the naïve and the Mandarin groups ($\chi^2(1) = 120.200$, $p < .0001$). As for LING, the F0 contours of responses to /a/ were significantly different from the F0 contours of responses to /ma1/ ($\chi^2(1) = 48.000$, $p < 0.0001$). As for MODE, the F0 contours in the AUDIO-ONLY condition were significantly different from the F0 contours in the AUDIO-VISUAL condition ($\chi^2(1) = 6.400$, $p < 0.05$). As for DIRECTION, the responses to upward shifts were significantly different from the responses to downward shifts ($\chi^2(1) = 24.200$, $p < 0.0001$). As for MAGNITUDE, the responses to pitch-shifted stimuli of 25 cents were significantly different from the responses to pitch-shifted stimuli of 200 cents ($\chi^2(1) = 33.000$, $p < 0.0001$).

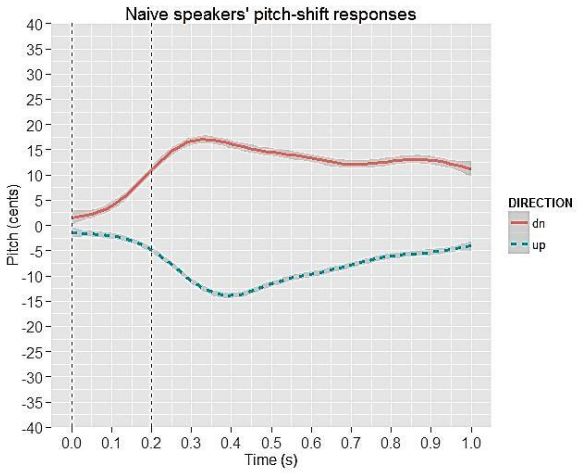
Table 5.2 Summary of final Generalized Additive Model in Study 3

Parametric coefficients				
	Estimate	Std. Error	*t*	*p*
intercept	11.020	0.891	12.366	<.0001
GROUPNaive	0.754	1.255	0.601	0.548
LINGMA	0.415	0.073	5.697	<.0001
MODEVISCOR	-1.557	0.073	-21.381	<.0001
DIRECTIONup	-2.254	0.073	-30.960	<.0001
MAGNITUDE200c	0.540	0.073	7.413	<.0001
Approximate significance of smooth terms				
	edf	Ref.*df*	*F*	*p*
s(SUBJ)	17.94	18.000	293.181	<.0001
s(TIME)	-3.509e-15	1.000	0.000	0.464
s(TIME):GROUPMandarin	1.045	1.122	0.144	0.733
s(TIME):GROUPNaive	6.517	7.450	4.577	<.0001
s(TIME):LINGAH	5.940	7.060	7.191	<.0001
s(TIME):LINGMA	0.600	0.600	2.109	0.261
s(TIME):MODEAUDIGN	6.628	7.656	13.876	<.0001
s(TIME):MODEVISCOR	0.600	0.600	0.399	0.625
s(TIME):DIRECTIONdn	7.350	7.697	5.944	<.0001
s(TIME):DIRECTIONup	6.367	6.829	2.315	<.05
s(TIME):MAGNITUDE025c	8.014	8.224	3.448	<.001
s(TIME):MAGNITUDE200c	2.791	2.996	0.324	0.807661

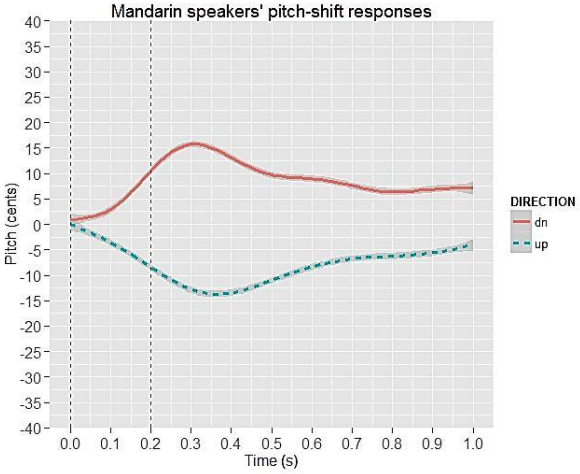
As the generalized additive models smoothed the F0 curves in temporal domain, the F0 contours resulting from the final model represent a smoothed curve in all cases. The estimated smoothed changes of the F0 contour in relation to the time dimension per condition built from the GAM are shown in Figure 5.4.

Figure 5.4 Estimated smoothed changes of the F0 contour in relation to the time dimension per condition built from the GAM. ***a.*** Naïve speakers' pitch-shift responses. ***b.*** Mandarin speakers' pitch-shift responses. ***c.*** Pitch-shift responses to /a/. ***d.*** Pitch-shift responses to /ma1/. ***e.*** Pitch-shift responses in the AUDIO-ONLY condition. ***f.*** Pitch-shift responses in the AUDIO-VISUAL condition. ***g.*** Pitch-shift responses to 25 cents perturbation. ***h.*** Pitch-shift responses to 200 cents perturbation. ***i.*** Pitch-shift responses to up-shifts and down-shifts.

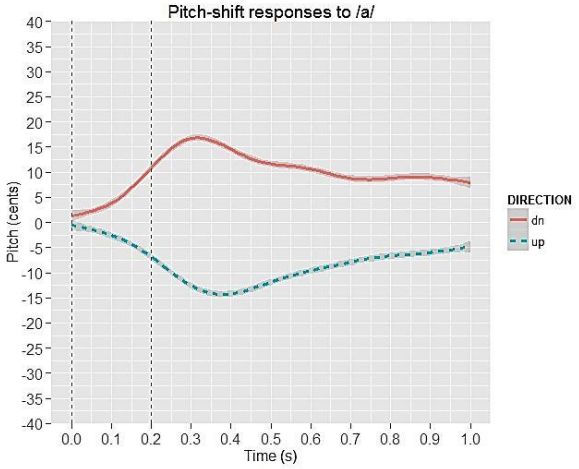
a.



b.



c.



d.

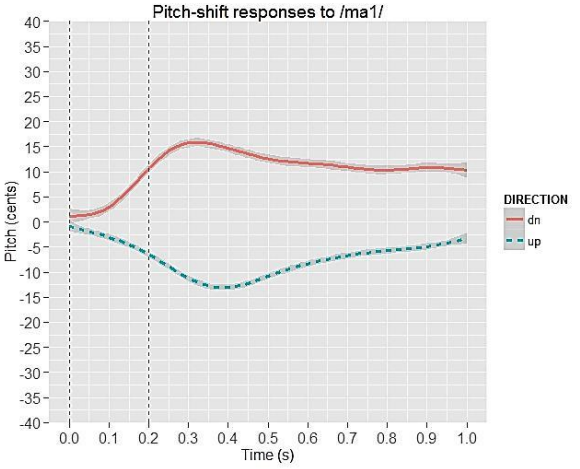
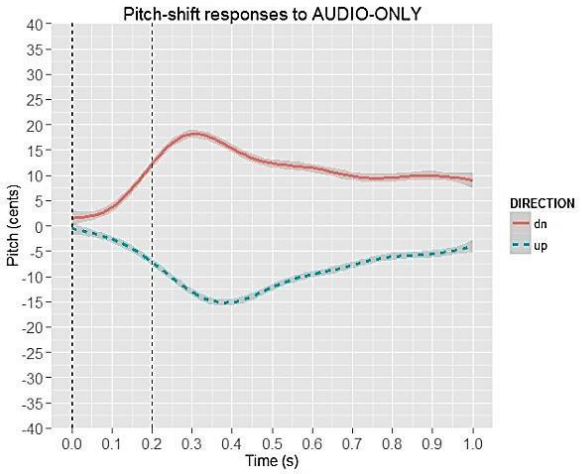
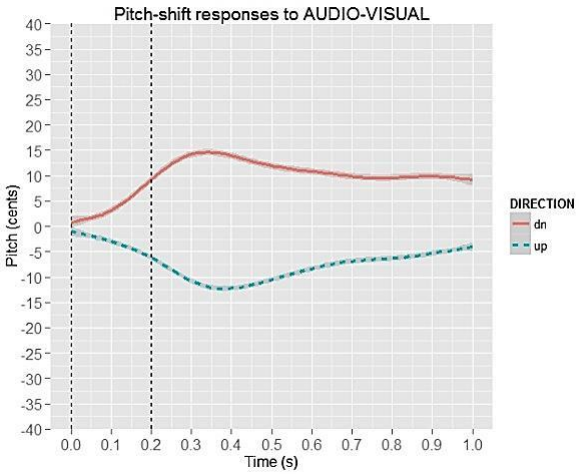


Figure 5.4 (Cont.)

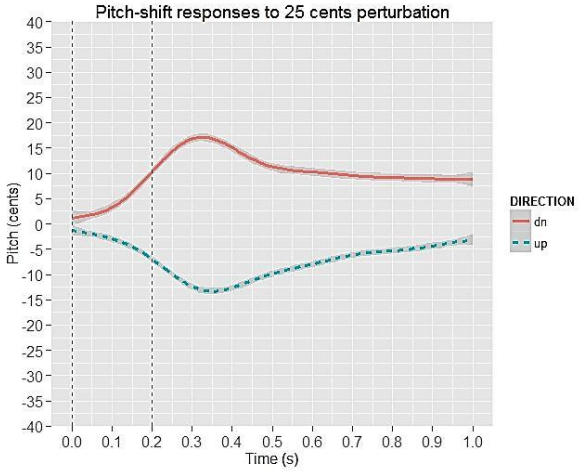
e.



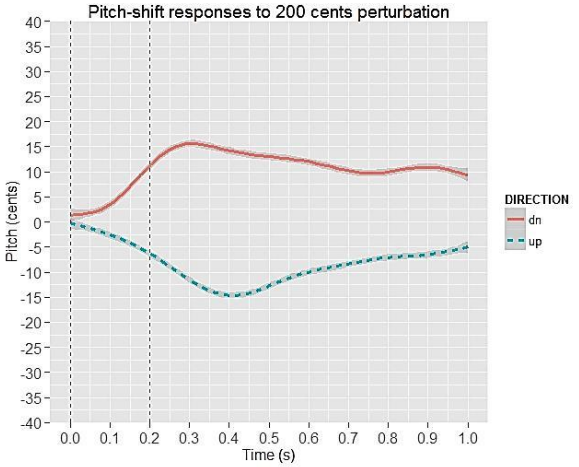
f.



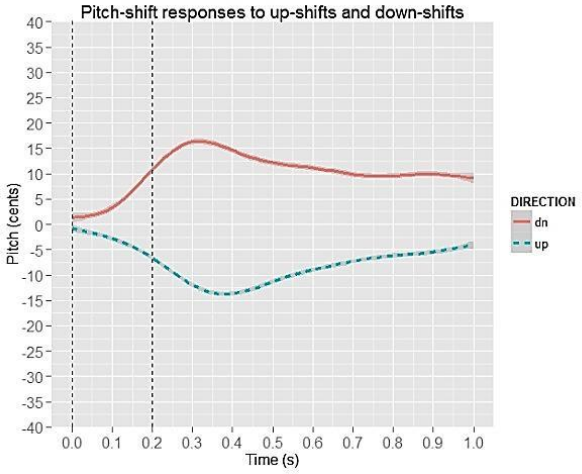
g.



h.



i.



Note. The x-axis is time measured in second. The y-axis represents the pitch-shift response

magnitude measured in cents. Red curves represent the responses to down-shift stimuli. Blue curves represent the responses to up-shift stimuli. The vertical dotted lines represent the onset and offset of the pitch-shift stimuli.

Note that the responses I included are compensatory responses, so the direction of responses is opposite to the direction of pitch-shift stimuli. The onset of pitch-shift stimuli is 0 second and the offset is 0.2 second. After the response peaks, Mandarin speakers quickly returned to the baseline (close to 0 cents) but naïve speakers did not, especially for the F0 contours of down-shift stimuli (GROUP effect). F0 contours of the downward shift /ma1/ did not quickly return to the baseline while the F0 contours of /a/ did (LING effect). Similarly, F0 contours in the 200 cents condition did not quickly return to the baseline, compared to the F0 contours in the 25 cents condition (MAGNITUDE effect). As found in the dependent variable *peak amplitude*, F0 contours in the AUDIO-VISUAL condition had larger amplitudes than the F0 contours in the AUDIO-ONLY condition (MODE effect). As for the DIRECTION effect, F0 contours of down-shift stimuli had shorter *peak times* than the F0 contours of up-shift stimuli, as shown in repeated measures ANOVA. Additionally, F0 contours of down-shift stimuli returned to approximately 10 cents after the response peaks, which was larger than the 5 cents in the F0 contours of up-shift stimuli.

5.4 Discussion

This study investigated whether real-time visual feedback facilitates pitch-shift response suppression in the contexts of simple vowels and tonal language syllables for Mandarin speakers and naïve speakers. The findings show that providing real-time visual feedback on voice F0 helps speakers to suppress pitch-shift responses (in both peak amplitude and relative gain). The advantage of using real-time visual feedback was observed in both Mandarin speakers and naïve speakers. However, Mandarin speakers had shorter peak times than naïve speakers, suggesting

that Mandarin speakers adjusted their voice F0 more rapidly in presence of pitch perturbation. Mandarin speakers also returned to the baseline (close to 0 cents) after the response peaks faster than naïve speakers. As for stimulus specificity, there was no significant difference between /a/ and /ma1/ in pitch-shift response latencies and pitch-shift response magnitudes. However, the F0 contours show it may be harder to stabilize the vocalization of Mandarin word /ma1/ than the sustained vowel /a/, because the F0 contours of the downward shift /ma1/ did not quickly return to the baseline after the response peaks. Pitch-shift responses to 25 cents perturbation had shorter peak times but larger relative gain than pitch-shift responses to 200 cents. The F0 contours show that it may be more difficult to stabilize the voice in response to 200 cents perturbation because the F0 contours in the 200 cents condition did not quickly return to the baseline as quickly after the response peaks. These results support that pitch-shift responses are tuned to correct small perturbations in voice F0. In addition, a DIRECTION effect on response latencies (onset time and peak time) and relative gain was also found, where responses to downward shifts had shorter onset times, shorter peak times, and larger relative gain than responses to upward shifts. The voice may be harder to stabilize in response to downward perturbation, because F0 contours of downward shifts only returned to approximately 10 cents after the response peaks, compared to 5 cents in the F0 contours of upward shifts. Regarding the findings for the nonlinguistic tone frequency discrimination task, no statistical differences between naïve speakers and Mandarin speakers were found. The tone discrimination ability was not correlated with any pitch-shift response measurements.

5.4.1 Suppression of Pitch-shift Responses with the Aid of Visual Feedback

The findings from the pitch-shift response amplitudes and relative gain show that providing speakers with real-time visual feedback for F0 leads to greater suppression of pitch-shift

responses. This corresponds to the evidence in balance control and golf learning which argues that providing visual feedback of external cues enhances motor skills related to perturbation suppression (Lauterbach et al., 1999; Peh et al., 2011; Shea & Wulf, 1999; Wulf et al., 1998; Wulf et al., 2010). In the current study, speakers' attention was directed either to auditory feedback changes in the AUDIO-ONLY condition (internal focus) or to real-time visual feedback for F0 tracking results in the AUDIO-VISUAL condition (external focus). When speakers' attention was directed to the visual F0 tracking results, they were more successful in suppressing their responses to pitch perturbation.

However, naïve speakers did not benefit from real-time visual feedback more than Mandarin speakers did. Both groups showed suppression effect in the AUDIO-VISUAL condition, irrespective of the stimuli (/a/ or /ma1/). This suggests that the advantage of visual feedback is not limited to a particular language group. All speakers can make use of the external feedback (real-time visual feedback) to stabilize or regulate their voice F0. Previous research has shown that visual feedback assists singers in pitch accuracy (Ferguson et al., 2005; Howard et al., 2007; Howard et al., 2004; Thorpe, 2002; Wilson et al., 2008) and L2 learners in tone/intonation production (Chun et al., 2013; Hardison, 2004; Levis & Pickering, 2004). The current research adds to evidence that visual feedback also improves the stability of voice F0 in presence of perturbation.

In the current study, the reference lines displayed in the real-time visual feedback had a range of mean F0 $\pm$5 Hz. The 5 Hz difference was equal to 57 cents for a male voice of 150 Hz and 43 cents for a female voice of 200 Hz. The 5 Hz difference in visual feedback is within a reasonable range for speakers to notice the changes in pitch and to respond by stabilizing their voice F0. It is not clear whether changing the reference line resolution would affect speakers' ability to regulate voice F0 or whether the 5 Hz difference is an optimal reference. This issue

should be explored further in the future. What can be argued for now with statistical support is that the combination of auditory and visual feedback is more effective for suppression of pitch-shift responses than auditory feedback only. This is promising in language learning. When learning new sounds, L2 learners first listen to model speech and then reproduce the sounds stored in memory. L2 learners may or may not be able to auditorily perceive the differences between their speech and model speech. Instantaneous visual feedback for the voice could enable L2 learners to better perceive and then correct vocal errors.

It is important to note that providing real-time visual feedback for F0 did not eliminate pitch-shift responses. Tye-Murray (1986) argued that presenting instant visual feedback of articulators did not completely diminish the effects of delayed auditory feedback (DAF), which, like pitch-shift paradigm, can be used to demonstrate the importance of auditory feedback for production (Chon, Kraft, Zhang, Loucks, & Ambrose, 2013; Kalinowski, Stuart, Sark, & Armson, 1996; Sparks, Grant, Milay, Walker-Baston, & Hynan, 2002; Van Borsel, Reunes, & Van de Bergh, 2003). The responses to pitch-shift stimuli or delayed stimuli are fast and not under voluntary control (longer latency voluntary responses were not elicited). The presence of pitch-shift responses in the AUDIO-VISUAL condition (though attenuated, compared to the AUDIO-ONLY condition) shows that reflex-like pitch-shift responses cannot be completely overridden by voluntary control over the voice (correcting voice F0 immediately when the F0 contours drifted above or below the reference lines). There is however, a range over which pitch shift responses are sensitive to task and language. Therefore, pitch-shift responses serve as a robust method to investigate how internal models are modulated for voice production.

5.4.2 Stability of Voice F0

Although naïve speakers and Mandarin speakers did not differ in pitch-shift response

latencies and peak amplitudes, they showed differences in the F0 contours after the response peaks. Mandarin speakers returned to the baseline faster than naïve speakers did. This GROUP difference suggests that Mandarin speakers are more likely to stabilize their voice using faster adjustment than naïve speakers. The fast adjustment in Mandarin speakers also occurs in the response *peak time*, where Mandarin speakers were faster than naïve speakers. Both findings suggest that Mandarin speakers have different internal models for tone and thus regulate voice F0 in a different manner.

F0 contours of the Mandarin word /ma1/ did not quickly return to the baseline, especially the responses to downward shifts, compared to the contours of sustained vowel /a/. This difference echoes the stimulus-specificity found in Study 2. Pitch-shift responses to the sustained vowel /a/ were suppressed in both Mandarin speakers and trained vocalists compared to naïve speakers, while the pitch-shift responses to Mandarin syllables were suppressed only in Mandarin speakers compared to naïve speakers. It is likely that controlling the pitch in a Mandarin word is different from that in a sustained vowel and requires language-specific training. Thus, stabilizing the voice F0 in Mandarin words could be more difficult than for linguistically simpler sounds.

Zarate et al. (2010) shows that singers were less able to suppress pitch-shift responses to 25 cents stimuli than to 200 cents stimuli, suggesting that pitch-shift responses to smaller shifts are under less voluntary control. The current study shows that responses to the 25 cents condition had larger relative gain (70.313 $\pm$ 2.770 %) than the responses to the 200 cents condition (9.128 $\pm$ 0.313 %). In addition, F0 contours in the 25 cents quickly returned to the baseline after the response peaks while the F0 contours in the 200 cents did not. All the findings suggest that responses to smaller pitch perturbation are indeed harder to suppress but they are still subject to adjustments. The results also support the claim that pitch-shift responses are tuned to the

correction of small pitch perturbation but are not invariant (Hain et al., 2000; Liu & Larson, 2007).

S. H. Chen et al. (2007) and Larson et al. (2001) argue that pitch-shift response latencies and amplitudes are not affected by the stimulus direction (upward or downward shifts) during sustained vowels. However, Liu, Meshman, Behroozmand, and Larson (2011) found that vocal response magnitudes were larger for down-shift stimuli than up-shift stimuli. The findings in the current study show that F0 contours of downward shifts did not return to the baseline after the response peaks (10 cents) as the F0 contours of upward shifts did (5 cents). This bears similarity to the DIRECTION effect found in Study 1, where in naïve speakers, the responses to down-shift stimuli (-50/-100 cents) showed more oscillation than responses to up-shift stimuli (+50/+100 cents). It is possible that stabilizing the voice in the presence of down-shift stimuli is more difficult. These suggest that regulating voice F0 in response to upward shifts and downward shifts depends on different mechanisms, though it remains unclear what the mechanisms exactly are. In the Mandarin condition, participants were told that the syllable /ma1/ has a high level tone and were instructed to hold the high level tone for 3 seconds. Their comfortable pitch for vocalizing /a/ was slightly lower than the /ma1/ (approximately 10 Hz lower). Both /ma1/ and /a/ seemed to be associated with high pitch for individual speakers. It is speculated that downward shifts could make the voice sound more deviant from the intended pitch and thus require more effort to adjust.

F0 contours after the response peaks give us more detail on how pitch-shift responses change over time. The whole contours tell us whether and how closely the responses return to the baseline. Therefore, I propose that, to account for the effects of GROUP, LING, MODE, DIRECTION, and MAGNITUDE in pitch-shift responses, F0 contours as well as response latencies and amplitudes should be considered.

5.5 Conclusion

The effect of real-time visual feedback on pitch-shift responses was examined in naïve speakers and Mandarin speakers. Both naïve speakers and Mandarin speakers benefited from the use of real-time visual feedback in that they showed increased suppression of pitch-shift responses (smaller response peak amplitudes and smaller relative gain) in the combined feedback condition (audio plus visual). This result suggests that directing speakers' attention to external feedback could improve the stability of their voice F0. GROUP differences were found for response latencies and F0 contours but not for response peak amplitudes. This finding extends some limited support to the idea that Mandarin speakers have more robust internal models for tone by being able to make more rapid F0 adjustments (faster peak time and faster return to the baseline) than naïve speakers. F0 contours in the /ma1/ condition did not quickly return to the baseline after the response peaks, suggesting that pitch-shift responses are stimulus-specific. It may be harder to stabilize the vocalization of Mandarin word /ma1/ than the sustained vowel /a/. F0 contours in the 25 cents condition quickly returned to the baseline after the response peaks, suggesting that pitch-shift responses are tuned to correct small perturbations.

CHAPTER 6
GENERAL DISCUSSION

6.1 The Main Issues

This dissertation addressed an important question in the field of second language acquisition: How does effective language learning reshape the underlying language system. Language learning becomes effortful and difficult especially after the critical period, because the underlying language system has been shaped in a certain way to understand linguistically meaningful contrasts in the first language. When second language learners (particularly adult learners) try to learn a new language, they have to reshape or reorganize their underlying language systems. This reshaping or reorganization process could occur at any linguistic levels including phonetic, phonological, syntactic or semantic representations. For instance, for native speakers of Japanese, the allophonic pair /r-l/ in Japanese has to be recategorized as separate phonemes when they learn English as a second language. Second language learning changes the way you encode and navigate linguistic representations. More importantly, the reshaping of internal representations which takes place in the brain determines whether a second language learner could achieve native-like language proficiency.

Linguists have been enthusiastic in investigating internal representations of language. Behavioral, ERP, and fMRI research have been used to explore how our brains process or store linguistic information. It has been shown that lexical tone perception is processed by the left hemisphere for native Mandarin speakers, but by the right hemisphere for non-tone speakers (Van Lancker & Fromkin, 1973; Y. Wang et al., 2001). This contrast suggests that the internal representations for tones are tuned by language experience. However, after a certain amount of tone training, the left hemisphere became activated for non-tone speakers (Yue Wang et al., 2003; Wong & Perrachione, 2007), suggesting that human brains are endowed with remarkable neural

plasticity. Nevertheless, tone learning involves not only perception but also production, which requires the acquisition of new motor skills. Controlling voice F0 in the syllable domain is a language-specific skill that Mandarin speakers acquire in early childhood. Learning new motor skills for voice F0 control relies on communication and coordination between the sensory system and the motor system. Internal models are thought to mediate sensory input and motor commands in order to achieve a certain movement goal. Therefore, the internal models are an essential gatekeeper that checks the accuracy of tone production.

This dissertation is the first research to examine development of the motor skills and reshaping of the internal models for tone in second language learners of Mandarin by using the pitch-shift paradigm. Responses to pitch perturbation have reflex-like properties and are hard to suppress even if speakers are notified of the change in voice F0. Results from the dissertation show that pitch-shift responses are affected by native language experience and extensive vocal training experience. Thus, the pitch-shift paradigm serves as a robust tool to investigate the stability of internal models for tone. Beyond the pitch-shift literature, the relationship between the ability to perceive tonal contrasts and the ability to correct pitch perturbation was examined. Ideally L2 learners should be able to perceive differences in tones in order to utter lexical tone accurately. It is likely speakers who have a stable internal model for tone would be good at correcting pitch perturbation as well as discriminating pitch differences.

This dissertation examined the language experience effect at a wider spectrum of language experience by asking whether naïve speakers who were never exposed to tonal languages, L2 learners of Mandarin, and native speakers of Mandarin would show a gradual change in their tone discrimination and pitch-shift responses. In addition to the language experience effect, the effect of vocal training experience was examined by asking whether trained vocalists have a potential to master tonal languages faster than nonsingers. Like native Mandarin speakers,

trained vocalists are supposed to be sensitive to pitch differences and have fine control over their voice F0 (though in a nonlinguistic domain). It would be interesting to see if the motor skills can be transferred from one domain to another.

Finally, my dissertation extended the pitch-shift literature by investigating a cross-modal advantage in voice F0 control that had never been studied. Although visual cues could create illusion in speech perception (e.g., McGurk effect (McGurk & MacDonald, 1976)), it has been shown that visual cues could enhance pitch accuracy in singers (Ferguson et al., 2005; Howard et al., 2007; Howard et al., 2004; Thorpe, 2002; Wilson et al., 2008) and tone/intonation production in L2 learners (Chun et al., 2013; Hardison, 2004; Levis & Pickering, 2004). This dissertation explored whether naïve speakers and Mandarin speakers can benefit from the use of visual feedback for F0 and exhibit suppressed pitch-shift responses.

In sum, this dissertation focused on the language experience and vocal training experience effects on vocal responses to pitch perturbation. By comparing different language groups, I hope to find compelling evidence for the reshaping of internal models for tone in second language learners of Mandarin. I also tried to find the relationship between perception and production by examining the ability to perceive tonal contrasts in both linguistic and nonlinguistic domains. At a practical end, whether real-time visual feedback is beneficial for stabilizing voice F0 was investigated. By directing speakers' attention to external cues, real-time visual feedback may help second language learners to learn new motor skills.

6.2 The Impact on Second Language Learning of Mandarin Tone

6.2.1 Pitch-shift Response as an Indicator of Language Proficiency

Traditionally, language proficiency has been evaluated by standardized tests. However, standardized tests measure students' performance instead of their competence. Results of

standardized tests may be subject to practice and thus be lack of credibility. We do not know what happened to a learner's brain after learning a second language. Pitch-shift responses are reflex-like responses and thus can be used to tap into the internal models for tone. It has potential to be developed as a good alternative for measuring language proficiency. Although continuous exposure to pitch perturbation (say 100 trials) causes an adaptation effect, no long-term adaptation to pitch perturbation has been reported (Jones & Munhall, 2000, 2002; Keough & Jones, 2009; Liu, Behroozmand, & Larson, 2010). As pitch-shift responses are not subject to practice, they can be used to assess a learner's capability of learning Mandarin tone and provide a valid evaluation.

From the results of Study 1-3, native-like (Mandarin-like) pitch-shift responses should be small in response peak amplitudes, fast in response peak time, and quickly returning to the baseline after the response peaks. This pattern suggests that Mandarin speakers are less affected by unexpected pitch perturbation and have fast adjustment for unstable voice F0. It is expected that advanced L2 learners would have the same characteristics in their pitch-shift responses as native Mandarin speakers. Our L2 learners' pitch-shift responses were not like those of native speakers of Mandarin, suggesting that the L2 learners have not acquired native-like internal models for tone (Study 1-2). However, the L2 learners' pitch-shift responses were not like naïve speakers' pitch-shift response either, suggesting that learning Mandarin has changed the way speakers control their voice F0. The group differences in pitch-shift responses lied not only in response latencies or response amplitudes but also in the whole F0 contours, suggesting that both point estimation and curvature estimation could serve as valid measurements for language proficiency. Due to the limited sample of L2 learners and the lack of L2 learners' proficiency scores, although pitch-shift responses provide an insight into language learning (native-like or not), it requires further research to understand what is a threshold for pitch-shift responses to

measure second language proficiency, and to justify the use of the pitch-shift responses for evaluating L2 learners' proficiency and capability.

Notice that pitch-shift responses are stimulus-specific. While both trained vocalists and Mandarin speakers could have suppressed pitch-shift responses to the sustained vowel /a/, only Mandarin speakers had suppressed pitch-shift responses to Mandarin high level tone (Study 2). Study 3 also shows that it may be harder to stabilize the vocalization of the Mandarin word /ma1/ than the sustained vowel /a/, because the F0 contours of the downward shift /ma1/ did not quickly return to the baseline after the response peaks. The findings from Study 2 and Study 3 suggest that producing Mandarin syllables is not the same as simple vowel vocalization and language-specific training is required. Thus, measuring second language proficiency with the use of pitch-shift responses requires more substantive work to establish the baseline of different types of test stimuli.

6.2.2 Interaction between Perception and Production in Language Learning

My hypothesis about the relation between perception and production is that audio-vocal response amplitude and latency are correlated with Mandarin tone discrimination ability. Study 2 shows that participants with more accurate Mandarin tone discrimination had smaller pitch-shift response ranges for tones. Tone discrimination ability is related to the correction for pitch perturbation, suggesting that language learning is a multidimensional construct (perception, production, and the interaction between perception and production). Learning Mandarin tone needs to blend the perception of new sensory information with new motor commands to achieve the desired result. The important connection between perception and production taking place in language learning can also be observed in bird song. For example, juvenile oscine songbirds first listen to and memorize the song of an adult male tutor. At the subsequent stage of vocal learning,

they match their own song to the memorized tutor model via auditory feedback. Successful song learning through early auditory exposure is crucial to the bird's reproduction. Absence of a tutor or being raised by another species will make the birds fail to attract females of their own kind (Doupe & Kuhl, 1999; Janik & Slater, 1997). Similarly in human speech, successful tone production depends on sufficient auditory exposure. Speakers' sensitivity to differences between sounds is related to their ability to control voice F0 (Study 2), suggesting that perception is essential in establishing internal models for tone. For native speakers of Mandarin who have been exposed to tone since early childhood, their perception was linguistically tuned and their language-specific motor commands were built for Mandarin tone production. For L2 learners to acquire native-like internal models for tone, they probably first need to be trained extensively in tone perception and discrimination. For trained vocalists who have been trained in pitch accuracy and voices, their sensitivity to pitch may facilitate the establishment of internal models for tone with focused training.

The dissertation also addressed whether tone discrimination ability is domain-specific (language domain or musical domain) by examining nonlinguistic tone discrimination and linguistic tone discrimination. It is possible that L2 learners who are good at differentiating musical tones may be more capable of learning Mandarin tone contrasts. Results from the three studies were not consistent. When the accuracy feedback was provided for each trial in the nonlinguistic tone discrimination task, performance in the nonlinguistic tone discrimination was correlated with the performance in the Mandarin tone discrimination task and Mandarin speakers were superior to naïve speakers (Study 1). However, when the accuracy feedback was covered, there was no significant relationship between the two perception tasks (Study 2) and no group differences (Study 2-3). It is possible that removing the accuracy feedback also eliminated potential learning effect. Tone perception still needs domain-specific training. Having good

hearing in musical frequency does not guarantee the success of linguistic tone discrimination. Having tonal language experiences does not appear to enhance the ability to discriminate musical pitch generally, although when feedback is provided, there could be variation in findings (as per study 1).

6.2.3 Musicality

Human vocalization includes singing, speaking, crying, laughing, etc. Singing and speaking are the most common types of vocal activities. Research has shown that musicians outperformed nonmusicians in lexical tone identification and discrimination (C.-Y. Lee & Hung, 2008). The reason why musical experience facilitated tone word learning for listeners without a tone language background could be that the advantage in the nonlinguistic domain can be transferred to the linguistic domain. It is possible that musicians' internal representations are similar in some way to tonal speakers'. In the adaptive Mandarin tone discrimination task of Study 2, trained vocalists were the only group that could achieve the highest level (which had different segments and different speakers' voices and required retrieval of the abstract tone representations) as Mandarin speakers did, even though trained vocalists did not advance as rapidly as Mandarin speakers. This confirms that musicians can make use of their perception skills and apply them to linguistic tone without further instruction.

In addition to tone discrimination, whether musicians' vocal skills can be applied to linguistic tone production without tone training was examined. Trained vocalists may apply their vocal skill in singing to linguistic tone production when they were instructed to imitate Mandarin tone. If musicians perform like Mandarin speakers, the internal models for musical pitch should be in common with the internal models for linguistic tone. If musicians do not perform like Mandarin speakers, then internal models should be domain-specific. Study 2 shows that trained

vocalists had suppressed pitch-shift responses as Mandarin speakers did when the stimuli were sustained vowels (/a/). However, when the stimuli were Mandarin syllables (especially for the high flat tone /ma1/), trained vocalists behaved differently from Mandarin speakers. This suggests that trained vocalists' internal models are established specifically for singing. The vocal skills for singing are not the same as the vocal skills for speech.

Although the internal models for musical pitch are not identical to the internal models for linguistic tone, it does not mean musical training is of no help for tone learning. Trained vocalists' performances in the pitch-shift task were closer to Mandarin speakers than L2 learners were. Trained vocalists had more suppression in the pitch-shift responses than L2 learners. Trained vocalists also showed their advantage in Mandarin tone discrimination. Therefore, it is expected that trained vocalists potentially can master the vocal skills for tone production faster than nonmusicians do, because of their sensitivity to pitch and fine control over voice. Musical training, especially vocal training, should facilitate second language learning of Mandarin tone.

6.2.4 External Feedback in Language Learning

Attention of external focus has been examined in balance control, golf putting, basketball shooting, and dart throwing (see Peh et al. (2011) for review), suggesting that use of an external focus for instruction may be beneficial during skill learning. In Study 3, external feedback was provided during vocalization. Participants not only "listened to" their voices but also "looked at" the outcome of vocalization. Their attention was directed to the visual feedback, the physical reference lines which served as an external focus. Participants were instructed to look at the pitch tracking results and to correct their voices immediately if the results drifted.

Learning to speak a language, like motion skill learning, requires feedback to enhance accuracy. The benefit of using external feedback has been attested in second language learning

(Chun et al., 2013; Hardison, 2004; Levis & Pickering, 2004). The results in Study 3 also support that providing speakers with real-time visual feedback for F0 could help them stabilize their voice F0 in presence of pitch perturbation. Stability is an essential feature in motor control. The more perturbation you can attenuate, the more stable your internal models are. Mandarin speakers have shown that they have robust and stable internal models for tone through suppressed pitch-shift responses (Study 1-2). We would expect advanced L2 learners who have well-established speech commands for tone to have native-like internal models, i.e., stable internal models for tone. As real-time visual feedback could enhance the stability of voice F0, it may also assist L2 learners to reshape internal models for tone to be as stable as possible.

External feedback provides an opportunity for learners to modify their performances instantly that traditional instruction (such as verbal instruction) may not offer. Listening to instructors' pronunciations or students' own pronunciations may not be enough for mastering Mandarin tone or acquiring native-like internal models. Learning with an aid of external feedback may become a mainstream in language acquisition.

6.2.5 Internal Models for Language Learning

The DIVA model proposed by Guenther (Guenther, 2006) argue that auditory feedback plays an important role in the production of new sounds. During babbling or early speech repetitions, new speech sounds are learned by storing an auditory target and using auditory feedback to guide speech motion. Repeated production tunes feedforward commands, which supplant the feedback-controlled signals. Therefore, in acquired sounds, speakers can produce the sounds without relying on the auditory feedback to adjust speech.

By looking at a speaker's response to pitch perturbation, we may speculate how the feedforward loop and feedback loop in internal models coordinate in voice F0 regulation.

Mandarin speakers have suppressed pitch-shift responses, irrespective of the stimulus type, suggesting that Mandarin speakers may have strong feedforward commands for tone in their internal models so that they could be less affected by the perturbation they heard from the auditory feedback. Adult Mandarin speakers' tone production, unlike infants' babbling which rely heavily on auditory feedback, have their feedforward commands outweigh the influence of auditory feedback. The weight of feedforward commands over feedback may contribute to the characteristics of stable internal models for tone in Mandarin speakers. For L2 learners and naïve speakers who do not have stable internal models for tone, they may still have to rely on auditory feedback to monitor their voice F0 in tone production.

As suggested by the DIVA model, the feedforward commands are established and strengthened via repeated practice in production. To reshape the internal models, L2 learners need continuous training in tone production. On the other hand, the feedforward commands should be domain-specific. Although trained vocalists bear some resemblance to Mandarin speakers, the feedforward commands for producing linguistic tone are still different from the feedforward commands for singing musical notes. Though having some vocal training may speed up the learning process, to establish internal models for tone, speakers need language-specific training on tone.

Internal models explain the language experience effect and vocal training effect in T11 and T12, but not in T14 where no group difference was found. It is likely that the feedforward motor command for falling tone is executed too fast for feedback to interfere with. While the feedforward command for falling tone is executed, it decreases the susceptibility to feedback. In other words, rapid changes in pitch may be controlled differently by a very strong feedforward motor command than sustained pitch, so the correction mechanism through the feedback system does not come into play.

This dissertation takes the initial steps to develop an understanding of internal models in language learning. The internal model approach provides an insight into how second language learning reshapes the underlying language system. L2 learners in Study 1 and Study 2 did not behave like naïve speakers, suggesting that the internal models of L2 learners are still changing. The L2 learners did not act like native Mandarin speakers either, suggesting that the internal models of L2 learners are not native-like. Pitch-shift responses may be an appropriate test of internal models in second language learning, though more data is needed to differentiate different proficiency levels of L2 learners. It is predicted that learners having native-like internal models for tone should be more capable of controlling Mandarin tone production. This dissertation diverges from other language learning research in a way that this dissertation looks at robust responses which give us a taste of the force and timing of neural commands for speech. This dissertation extends the study of language learning evaluation to the study of language learning in capability. It is believed that internal model or underlying language system directly reflects the state of tone learning and can be developed to evaluate language proficiency and to estimate tone learning aptitude.

6.3 Future Research

The effects of language experience and musical experience on second language tone learning have been examined from the perspective of internal models. How external feedback may facilitate regulation of voice F0 was also investigated. Results have shown that L2 learners are different than Mandarin speakers in both tone perception and tone production. What remains unclear is whether internal models for tone would reveal developmental changes if we investigate L2 learners with different proficiency levels. The students recruited from the campus were beginners and had remarkable variance in their performances even though they enrolled in

the same level of Chinese class. It would be interesting to compare L2 learners of Mandarin who have learned Mandarin for a longer period of time (say 10 years) with those who just started learning Mandarin.

Trained vocalists' pitch-shift responses and their tone discrimination were explored in the dissertation. It is unclear whether other musicians mastering in various instruments instead of voice would have the same advantage in both tone production and perception as trained vocalists do. Musicians mastering in instruments may have good pitch discrimination but may not necessarily have good control of voice F0. It would be interesting to see what kind of musical experience can facilitate second language learning of Mandarin tone.

Study 3 examined whether real-time visual feedback could help speakers stabilize voice F0. The external feedback may be of help for L2 learners to reshape their internal models to be as stable as Mandarin speakers. Ultimately, we hope to see that L2 learners can behave like Mandarin speakers without the help of real-time visual feedback. However, it is still unclear how much training with visual feedback L2 learners would need and whether the advantage of real-time visual feedback has a long-term effect on tone production.

Pitch-shift responses could serve as an indicator for language proficiency. From the discriminant analyses and generalized additive model analyses, we know that pitch-shift responses are an important predictor for a speaker's language background. However, for pitch-shift responses to become standardized measurements, we will need large samples to justify the features in each language group and in each level of proficiency.

Finally, feedforward commands in the internal models are built via repeated practice. The feedforward commands are domain-specific because singing in tune is different from speaking in tune. What remains interesting is whether feedforward commands are language-specific. Tone speakers may have an advantage for learning another tonal language. The advantage has been

shown in tone perception (Francis et al., 2008; Wayland & Guion, 2004). It would be interesting to see if tone speakers with different native tonal languages would have the same pitch-shift responses when they are instructed to produce lexical tones.

CHAPTER 7
CONCLUSIONS

This dissertation brings together diverse methodologies, including perception and production experiments, to investigate the reshaping of internalized pitch representations in the brain. It fills the current knowledge gap about whether the mechanisms for vocal control in speakers without a tonal language background can be reformed by language learning. The use of pitch-shift paradigm contributes to a more unified understanding of language learning and the underlying changes in the language system. The focus on audio-vocal interactions extends *qualitative-descriptive* approaches to describe language learning ability and turns them into *quantitative-predicative* models that can be developed for language assessment and aptitude testing. By assessing contributions of auditory feedback and real-time visual feedback, this dissertation opens new lines of research into cross-modal sensory influences on language learning.

In Study 1, I sought baseline information regarding whether pitch-shift responses to the vowel /a/, non-linguistic tone discrimination, and linguistic tone discrimination would differ in naïve adults (no exposure to tonal languages), L2 adult learners of Mandarin, and native Mandarin speakers. Clear advantages were found for tone discrimination in Mandarin speakers and L2 adult learners. Mandarin speakers were less affected by the magnitude of pitch perturbation, showing that their internal models for tone are more stable or robust than naïve speakers and L2 learners. Study 2 built on this finding by looking at additional perception-production factors and incorporating musicianship. Study 2 measured pitch-shift responses to actual Mandarin tone sequences and assessed tone perception learning using an adaptive test of tone discrimination. Results show that certain advantages for tone perception

were associated with the ability to correct for pitch-shift in production. Although vocal training may potentially facilitate tone learning, mastery of tone production still requires language-specific training. Study 2 also found stimulus-specific pitch-shift responses: Mandarin speakers had larger suppression in tone 1 and tone 2 than in tone 4, compared to naïve speakers who showed robust compensatory pitch-shift responses. Trained vocalists resembled Mandarin speakers by displaying suppressed pitch-shift responses that were not seen in L2 learners. Study 3 explored how enhancing feedback by adding real-time visual feedback to tone production could influence suprasegmental control critical for tones. Results show that both naïve speakers and Mandarin speakers can benefit from the use of real-time visual feedback for stabilizing their voice F0.

This sequence of studies reveals important relations between auditory feedback, voice production and perception in naïve speakers, L2 learners and native speakers. By introducing a fourth group of naïve speakers with distinct musicianship, the role of 'audiomotor' factors not specific to language could be assessed. This dissertation constitutes an ambitious step in uncovering how specific and general experiences with auditory feedback impact language learning. This information opens the door to further studies of potentially malleable neural mechanisms involved in tonal language acquisition.

The U.S. Government considers Mandarin language learning to be crucial to U.S. national security, as evidenced by the National Security Education Program (NSEP). This dissertation contributes to the success of this national initiative by focusing on tonal language testing, teaching and training. L2 testing is mostly designed for examining learning outcomes, but outcomes are affected by both internal factors, such as internalized pitch representations in the brain, and external factors, such as language experience, musical experience, and instruction. The contribution of internal factors is often marginalized in language testing and learning.

Examiners and learners may be unaware of the neural plasticity that is required for processing lexical tone. Understanding internalized pitch representations in the brain provides a way to predict capability for learning Mandarin tones. Examining whether internalized pitch representations are reshaped will complement traditional language proficiency testing. Additionally, traditional teacher-based feedback and examples of production may have limitations. Providing specific real-time visual feedback that prompts suprasegmental voice changes can accelerate L2 tone learning. Establishing the cross-modal auditory-visual feedback influences provides pedagogical innovations for language instructors and L2 learners.

REFERENCES

Abbs, J. H., Gracco, V. L., & Cole, K. J. (1984). Control of multi-movement coordination: Sensorimotor mechanisms in speech motor programming. *Journal of Motor Behavior, 16*, 195-232.

Bauer, J. J., & Larson, C. R. (2003). Audio-vocal responses to repetitive pitch-shift stimulation during a substained vocalization: Improvements in methodology for the pitch-shifting technique. *Journal of the Acoustical Society of America, 114*(2), 1048-1054.

Behroozmand, R., & Larson, C. R. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neuroscience, 12*(54).

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of Acoustic Society of America, 109*(2), 775-794.

Bidelman, G. M., Gandour, J., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience, 23*(2), 425-434.

Binnie, C. A., Daniloff, R. G., & Buckingham, H. W. (1982). Phonetic disintegration in a five-year old following sudden hearing loss. *Journal of Speech and Hearing Disorders, 47*, 181-189.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods, 5*, 362-387.

Bolinger, D. (1978). Intonation across languages. In J. H. G. C. A. Ferguson & E. A. Moravcsik (Eds.), *Universals of Human Language, Vol. 2: Phonology* (pp. 471-524). Standford, CA: Stanford University Press.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Haskins Laboratories Status Report on Speech Research, SR-99/100*, 69-101.

Burke, J. (1996). Connections. *Scientific American, 274*(5), 110-111.

Burnett, T. A., Freeland, M. B., & Larson, C. R. (1998). Voice F0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America, 103*(6), 3153-3161.

Burnett, T. A., & Larson, C. R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *Journal of the Acoustical Society of America, 112*(3), 1058-1063.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Adaptive auditory feedback control of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. *Journal of Acoustic Society of America, 128*(4), 2033-2048.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage, 22*, 1182-1194.

Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *Journal of Acoustic Society of America, 124*(3), 1638-1652.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007a). Experience-dependent neural plasticity is sensitive to shape of pitch contours. *NeuroReport, 18*(18), 1963-1967.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007b). Mismatch negativity to pitch contours is influenced by language experience. *Brain Research, 1128*, 148-156.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2009a). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language, 108*, 1-9.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2009b). Sensory processing of linguistic pitch as reflected by the mismatch negativity. *Ear & Hearing, 30*(5), 552-558.

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences of the United States of America, 110*(7), 2653-2658.

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology Academia Sinica, 4*(2), 121-134.

Chen, S. H., Liu, H., Xu, Y., & Larson, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *Journal of the Acoustical Society of America, 121*(2), 1157-1163.

Chen, Z., Liu, P., Wang, E. Q., Larson, C. R., Huang, D., & Liu, H. (2012). ERP correlates of language-specific processing of auditory pitch feedback during self-vocalization. *Brain and Language, 121*, 25-34.

Cheyne, H. A., Kalgaonkar, K., Clements, M., & Zurek, P. (2009). Talker-to-listener distance effects on speech production and perception. *The Journal of the Acoustical of Society of America, 126*(4), 2052-2060.

Ching, Y. C. (1984). Lexical tone pattern learning in Cantonese children. *Language Learning and Communication, 3*(3), 317-334.

Chomphan, S., & Chompunth, C. (2012). Fujisaki's model of Thai's fundamental frequency contours with environmental noises. *American Journal of Applied Sciences, 9*(8), 1251-1258.

Chon, H., Kraft, S. J., Zhang, J., Loucks, T., & Ambrose, N. G. (2013). Individual variability in delayed auditory feedback effects on speech fluency and rate in normally fluent adults. *Journal of Speech, Language, and Hearing Research, 56*, 489-504.

Chun, D. M. (1989). Teaching tone and intonation with microcomputers. *The Computer Assisted Language Instruction Consortium Journal, 7*(1), 21-46.

Chun, D. M., Jiang, Y., & Avila, N. (2013). *Visualization of tone for learning Mandarin Chinese.* Paper presented at the Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference, Ames, IA: Iowa State University.

Ciocca, V., & Lui, J. Y. K. (2003). The development of the perception of Cantonese lexical tones. *Journal of Multilingual Communication Disorders, 1*(2), 141-147.

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *Journal of the Acoustical Society of America, 131*(6), 4756-4769.

Cowie, R., & Douglas-Cowie, E. (1992). *Postlingually acquired deafness*. New York: Mouton de Gruyter.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

Deutsch, D., Henthorn, T., & Dolson, M. (2000). Bilingual speakers perceive a musical illusion in accordance with their first language. *Journal of the Acoustical Society of America, 108*, 2591.

Donath, T. M., Natke, U., & Kalveram, K. T. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *Journal of the Acoustical Society of America, 111*(1), 357-366.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience, 22*, 567-631.

Duanmu, S. (1994). Against contour tone units. *Linguistic Inquiry, 25*(4), 555-608.

Eldridge, M., Saltzman, E., & Lahav, A. (2010). Seeing what you hear: Visual feedback improves pitch recognition. *European Journal of Cognitive Psychology, 22*(7), 1078-1091.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature, 453*(7198), 1102-1106.

Ferguson, S., Moere, A. V., & Cabrera, D. (2005). Seeing sound: Real-time sound visualization in visual feedback loops used for training musicians. *IEEE Proceedings of the International Conference on Information Visualisation*, 97-102.

Flanagan, J. (1972). *Speech Analysis Synthesis and Perception*. New York: Springer-Verlag.

Flanagan, J. R., & Wing, A. M. (1997). The role of internal models in motion planning and control: Evidence from grip force adjustments during movements of hand-held loads. *The Journal of Neuroscience, 17*(4), 1519-1528.

Flege, J. E. (1995). Second language speech learning theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-277). Timonium, MD: York Press.

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics, 36*, 268-294.

Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of Acoustical Society of Japan, 5*(4), 233-242.

Fujisaki, H., Ohno, S., & Gu, W. (2004). Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their F0 contours. *Proceedings of International Symposium on Tonal Aspects of Languages with Emphasis on Tone Language*, 61-64.

Fujisaki, H., Wang, C., Ohno, S., & Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model. *Speech Communication, 47*, 59-70.

Goldsmith, J. (1994). Tone languages. In E. Asher & J. Simpson (Eds.), *Encyclopedia of Language and Linguistics* (pp. 4626-4628). New York: Elsevier Science.

Gu, W., Hirose, K., & Fujisaki, H. (2007). Analysis of tones in Cantonese speech based on the command-response model. *Phonetica, 64*, 29-62.

Guenther, F. H. (1995). Speech sound acquisition coarticulation and rate effects in a neural network model of speech production. *Psychological Review, 102*, 594-621.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39*(5), 350-365.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural Modeling and Imaging of the Cortical Interactions underlying syllable production. *Brain and Language, 96*, 280-301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigateion of reference frames for the planning of speech movements. *Psychological Review, 105*, 611-633.

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Experimental Brain Research, 130*, 133-141.

Hain, T. C., Burnett, T. A., Larson, C. R., & Kiran, S. (2001). Effects of delayed auditory feedback (DAF) on the pitch-shift reflex. *Journal of the Acoustical Society of America, 109*(5), 2146-2152.

Halle, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs French listeners. *Journal of Phonetics, 32*, 395-421.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Techonology, 8*(1), 34-52.

Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M., & Ford, J. M. (2005). Fine-tuning of auditory cortex during speech production. *Psychophysiology, 42*, 180-190.

Held, R., & Hein, A. V. (1958). Adaptation of disarranged hand-eye coordination contigent upon re-afferent stimulation. *Perceptual and Motor Skills, 8*(87-90).

Henthorn, T., & Deutsch, D. (2007). Ethnicity versus early environment: Comment on 'Early Childhood Music Education and Predisposition to Absolute Pitch: Teasing Apart Genes and Evironment' by Peter K. Gregersen, Elena Kowalsky, Nina Kohn, and Elizabeth West Marvin [2000]. *American Journal of Medical Genetics Part A 143 A*, 102-103.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69*(3), 407-422.

Hirschberg, J. (1988). *Controlling intonational meaning in synthetic speech: Varying stress, pitch, and timing to convey discouse information*. Paper presented at the Columbia University Department of Computer Science AI Day.

Ho, C. S.-H., & Bryant, P. (1997). Development of Phonological Awareness of Chinese Children in Hong Kong. *Journal of Psycholinguistic Research, 26*(109-126).

Ho, C. S.-H., & Chan, D. W.-O. (2002). The cognitive profile and multiple-deficit hypothesis in Chinese developmental dyslexia. *Developmental Psychology, 38*(4), 543-553.

Ho, C. S.-H., Chan, D. W.-O., Lee, S.-H., Tsang, S.-M., & Luan, V. H. (2004). Cognitive profiling and preliminary subtyping in Chinese developmental dyslexia. *Cognition, 91*(1), 43-75.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*, 1213-1216.

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research, 45*, 295-310.

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125-1138.

Howard, D. M., Brereton, J., Welch, G. F., Himonides, E., DeCosta, M., Williams, J., & Howard, A. W. (2007). Are real-time displays of benefit in the singing studio? An exploratory study. *Journal of Voice, 21*(1), 20-34.

Howard, D. M., Welch, G. F., Brereton, J., Himonides, E., DeCosta, M., Williams, J., & Howard, A. W. (2004). WinSingad: A real-time display for the singing studio. *Logopedics Phoniatrics Vocology, 29*(3), 135-144.

Janik, V. M., & Slater, P. J. B. (1997). Vocal learning in mammals. *Advanced in the Study of Behavior, 26*, 59-99.

Jones, J. A., & Keough, D. (2008). Auditory-motor mapping for pitch control in singers and nonsingers. *Experimental Brain Research, 190*, 279-287.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback. *Journal of the Acoustical Society of America, 108*(3), 1246-1251.

Jones, J. A., & Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics, 30*, 303-320.

Jones, J. A., & Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Current Biology, 15*, 1768-1772.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models Supervised learning with a distal teacher. *Cognitive Science, 16*(307-354).

Kaan, E., Barkley, C. M., Bao, M., & Wayland, R. (2008). Thai lexical tone perception in native speakers of Thai, English and Mandarin Chinese: An event-related potentials training study. *BMC Neuroscience, 9*(1), 53-69.

Kalinowski, J., Stuart, A., Sark, S., & Armson, J. (1996). Stuttering amelioration at various auditory feedback delays and speech rates. *European Journal of Disorders of Communication, 31*, 259-269.

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *NeuroReport, 20*(14), 1231-1234.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology, 9*(6), 718-727.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Fumctionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 812-832.

Keough, D., & Jones, J. A. (2009). The sensitivity of auditory-motor representations to subtle changes in auditory feedback while singing. *Journal of Acoustic Society of America, 126*(2), 837-846.

Koelsch, S., Schroger, E., & Tervaniemi, M. (1999). Superior attentive and pre-attentive auditory processing in musicians. *NeuroReport, 10*, 1309-1313.

Kollmeier, B., Brand, T., & Meyer, B. (2008). Perception of Speech and Sound. In J. Benesty, M. M. Sondhi & Y. Huang (Eds.), *Springer handbook of speech processing* (pp. 65): Springer.

Kosling, K., Kunter, G., Baayen, H., & Plag, I. (2013). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech*. doi: 10.1177/0023830913478914

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research, 25,* 161-168.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608.

Lalazar, H., & Vaadia, E. (2008). Neural basis of sensorimotor learning Modifying internal models. *Current Opinion in Neurobiology, 18,* 573-581.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14,* 677-709.

Lane, H., & Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *Journal of the Acoustical Society of America, 89,* 859-866.

Larson, C. R. (1998). Cross-modality influences in speech motor control The use of pitch shifting for the study of F0 control. *Journal of Communication Disorders, 31,* 489-503.

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., & Hain, T. C. (2001). Comparison of voice F0 responses to pitch-shift onset and offset conditions. *Journal of the Acoustical Society of America, 110*(6), 2845-2848.

Larson, C. R., Burnett, T. A., Kiran, S., & Hain, T. C. (2000). Effects of pitch-shift velocity on voice F0 responses. *Journal of the Acoustical Society of America, 107*(1), 559-564.

Lauterbach, B., Toole, T., & Wulf, G. (1999). The learning advantages of an external focus of attention in golf. *Research Quarterly for Exercise and Sport, 70*(2), 120-126.

Lee, C.-Y., & Hung, T.-H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *Journal of Acoustic Society of America, 124*(5), 3235-3248.

Lee, C.-Y., & Lee, Y.-F. (2010). Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *Journal of Acoustic Society of America, 127*(1), 481-490.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech, 53*(2), 217-243.

Lee, D. N., & Aronson, E. (1974). Visual proprioceptive control of standing in human infants. *Perception & Psychophysics, 15*(3), 529-532.

Lee, Y.-S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A Cross-linguistic comparison. *Journal of Psycholinguistic Research, 25*(5), 527-542.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York: Wiley.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-75.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System, 32*, 505-524.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarinspeaking children. *Journal of Child Language, 4*(2), 185-199.

Li, W.-S., & Ho, C. S.-H. (2011). Lexical tone awareness among Chinese children with developmental dyslexia. *Journal of Child Language, 38*, 793-808.

Lin, C. Y., Wang, M., & Shu, H. (2012). The processing of lexical tones by young Chinese children. *Journal of Child Language, FirstViewArticles*, 1-15.

Liu, H., Auger, J., & Larson, C. R. (2009). Voice fundamental frequency modulates vocal response to pitch perturbations during English speech. *Journal of Acoustical Society of America, 127*(1).

Liu, H., Behroozmand, R., & Larson, C. R. (2010). Enhanced neural responses to self-triggered voice pitch feedback perturbations. *NeuroReport, 21*(7), 527-531.

Liu, H., & Larson, C. R. (2007). Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex. *Journal of the Acoustical Society of America, 122*(6), 3671-3677.

Liu, H., Meshman, M., Behroozmand, R., & Larson, C. R. (2011). Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback. *Clinical Neurophysiology, 122*, 951-957.

Liu, H., Wang, E. Q., Chen, Z., Liu, P., Larson, C. R., & Huang, D. (2010). Effects of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations. *Journal of the Acoustical Society of America, 128*(6), 3739-3746.

Liu, H., Xu, Y., & Larson, C. R. (2009). Attenuation of vocal responses to pitch perturbations during Mandarin speech. *Journal of the Acoustical Society of America, 125*(4), 2299-2306.

Maddieson, I. (1978). Universals of tone. In J. H. Greenberg (Ed.), *Universals of Human Language* (Vol. 2, pp. 335-366). Standford: Standford University Press.

Mandell, J. (2009). Adaptive Pitch Test: Accurately Measure Your Pitch Perception Abilities, from http://tonometric.com/adaptivepitch/

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*(3), 241-265.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition, 106*, 1367-1381.

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., & Wallace, M. E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary Issues in Communication Science and Disorders, 31*, 105-122.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Mitsuya, T., Samson, F., Menard, L., & Munhall, K. G. (2013). Language dependent vowel representation in speech production. *Journal of the Acoustical Society of America, 133*(5), 2993-3003.

Natke, U., Donath, T. M., & Kalveram, K. T. (2003). Control of voice fundamental frequency in speaking versus singing. *Journal of Acoustic Society of America, 113*(3), 1587-1593.

Newman, P. (1986). Contour tones as phonemic primes in Grebo. In K. Bogers, H. v. d. Hulst & M. Mous (Eds.), *The Phonological Representation of Suprasegmentals: Studies on African Languages* (pp. 175-193). Dordrecht: Foris Publications.

Odden, D. (1995). Tone: African languages. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 445-475). Cambridge: Blackwell.

Ohala, J. J., & Ewan, W. G. (1973). Speed of pitch change. *Journal of Acoustic Society of America, 53*(1), 345-345.

Parkinson, A. L., Flagmeier, S. G., Manes, J. L., Larson, C. R., Rogers, B., & Robin, D. A. (2012). Understanding the neural mechanisims invovled in sensory control of voice production. *NeuroImage, 61*, 314-322.

Peh, S. Y.-C., Chow, J. Y., & Davids, K. (2011). Focus of attention and its impact on movement behavior. *Journal of Science and Medicine in Sport, 14*, 70-78.

Pelegrin-Garcia, D., Smits, B., Brunskog, J., & Jeong, C.-H. (2011). Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical of Society of America, 129*(4), 1981-1990.

Penfield, W. (1959). *Speech and Brain Mechanisms*. New York: Atheneum.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speechh motor control Acoustic goals saturation effects auditory feedback and internal models. *Speech Communication, 22*, 227-250.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* PhD dissertation, Massachusetts Institute of Techonology.

Pike, K. L. (1948). Tone Languages: A technique for determing the number and type of pitch contrasts in a language, with studies in tonemic substitution and fusion. *University of Michigan Publications in Linguistics v. 4*.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*, 107-142.

Sakai, S. (2004). Additive modeling of English F0 contour for speech synthesis. *Proceedings of Internation Conference on Acoustics, Speech, and Signal Processing, 1*, 277-280.

Scheerer, N. E., & Jones, J. A. (2012). The relationship between vocal accuracy and variability to the level of compensation to altered auditory feedback. *Neuroscience Letters, 529*(2), 128-132.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 209-231.

Shea, C. H., & Wulf, G. (1999). Enhancing motor learning through external-focus instructions and feedback. *Human Movement Science, 18*, 553-571.

Shih, C. (1997). Intonation: Theory, models and applications. *Proceedings of an ESCA Workshop*, 293-296.

Shih, C. (2000). A Declination Model of Mandarin Chinese. In A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology* (pp. 243-268). The Netherlands: Kluwer Academic Publishers.

Shih, C., & Ao, B. (1997). Duration study for the Bell Laboratories Mandarin Text-to-Speech System. In J. v. Santen, R. Sproat, J. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis* (pp. 382-399). New York: Springer-Verlag.

Shih, C., & Lu, H.-Y. D. (2010). *Prosody transfer and suppression: Stages of tone acquisition.* Paper presented at the Speech Prosody, Chicago.

Shih, C., Lu, H.-Y. D., Sun, L., Huang, J.-T., & Packard, J. (2010). *An adaptive training program for tone acquisition.* Paper presented at the Speech Prosody, Chicago.

Siok, W. T., & Fletcher, P. (2001). The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. *Developmental Psychology, 37*(6), 886-899.

Sparks, G., Grant, D., Milay, K., Walker-Baston, D., & Hynan, L. (2002). The effect of fast speech rate on stuttering frequency during delayed auditory feedback. *Journal of Fluency Disorders, 27*, 187-201.

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schroger, E. (2005). Pitch discrimination accuracy in musicans vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research, 161*, 1-10.

Thorpe, C. W. (2002). Visual feedback of acoustic voice features in voice training. *Proceedings of the 9th Australian International Conference on Speech Science & Technology*, 349-354.

Tourville, J., Guenther, F. H., Ghosh, S., Reilly, K., Bohland, J., & Nieto-Castanon, A. (2005). Effects of acoustic and articulatory perturbation on cortical activity uring speech production. *Proceedings of the 11th Annual Meeting of the Organization for Human Brain Mapping, 26*(S1), S49.

Tse, J. K.-P. (1978). Tone acquisition in Cantonese: a longitudinal case study. *Journal of Child Language, 5*(2), 191-204.

Tye-Murray, N. (1986). Are real-time displays of benefit in the singing studio? An exploratory study. *Journal of Acoustic Society of America, 79*(4), 1169-1171.

Van Borsel, J., Reunes, G., & Van de Bergh, N. (2003). Delayed auditory feedback in the treatment of stuttering: Clients as consumers. *International Journal of Language & Communication Disorders, 38*, 119-129.

Van Lancker, D., & Fromkin, V. A. (1973). Cerebral dominance for pitch contrasts in tone language speakers and in musically untrained and trained English speakers. *Journal of Phonetics, 6*, 19-23.

Ventura, M. I., Nagarajan, S. S., & Houde, J. F. (2002). Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience, 10*(58), 58.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of Acoustic Society of America, 122*(4), 2306-2319.

Waldstein, R. (1990). Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *Journal of the Acoustical Society of America, 88*, 2099-2114.

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain and Language, 78*, 332-348.

Wang, Y., Sereno, J. A., Jongman, A., & Hirsch, J. (2003). fMRI evidence for cortical modification during learning of Mandarin lexical tone. *Journal of Cognitive Neuroscience, 15*(7), 1019-1027.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of Acoustic Society of America, 106*(6), 3649-3658.

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning, 54*(4), 681-712.

Welch, G. F. (1985). A schema theory of how children learn to sing in tune. *Psychology of music, 13*(1), 3-18.

Weltens, B., & Bot, K. d. (1984). Visual feedback of intonationa II: Feedback delay and quality of feedback. *Language and Speech, 27*, 79-88.

Wilson, P. H., Lee, K., Callaghan, J., & Thorpe, C. W. (2008). Learning to sing in tune Does real time visual feedback help. *Journal of Interdisciplinary Music Studies, 2*(1&2), 157-172.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*, 565-585.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B), 73*(1), 3-36.

Wulf, G., Höß, M., & Prinz, W. (1998). Instructions for motor learning: Differential effects of internal versus external focus of attention. *Journal of Motor Behavior, 30*(2), 169-179.

Wulf, G., Shea, C. H., & Lewthwaite, R. (2010). Motor skill learning and performance: A review of influential factors. *Medical Education, 44*, 75-84.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61-83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics, 27*, 55-105.

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of the Acoustical Society of America, 120*(2), 1063-1074.

Xu, Y., Krishnan, A., & Gandour, J. T. (2006). Specificity of experience-dependent pitch representation in the brainstem. *NeuroReport, 17*(15), 1601-1605.

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *Journal of the Acoustical Society of America, 116*(2), 1168-1178.

Yip, M. (1980). *The Tonal Phonology of Chinese*. Bloomington, Indiana: Indiana University Linguistics Club.

Yip, M. (2002). *Tone*. New York: Cambridge University Press.

Zarate, J. M., Wood, S., & Zatorre, R. J. (2010). Neural networks involved in voluntary and involuntary vocal pitch regulation in experienced singers. *Neuropsychologia, 48*, 607-618.

Zarate, J. M., & Zatorre, R. J. (2005). Neural substrates governing audio-vocal integration for vocal pitch regulation in singing. *Annals of the New York Academy of Sciences, 1060*(1), 404-408.

Zarate, J. M., & Zatorre, R. J. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. *NeuroImage, 40*, 1871-1887.